

Silicon

P. Siffert E.F. Krimmel (Eds.)

Silicon

Evolution and Future of a Technology

With 288 Figures and 21 Tables

P. Siffert

General Secretary European Materials Research Society
23 rue du Loess, 67037 Strasbourg Cedex 2, France

E.F. Krimmel

Siemens Corporate Research Laboratories, München, Germany (retired)
and
J.W. Goethe University, Frankfurt/M, Germany (retired)
and
European Materials Research Society, Strasbourg
and
Mendelssohnstr. 7, 82049 Pullach/Isartal, Germany

Dedicated to the European Materials Research Society

Library of Congress control Number: 2004102075

ISBN 978-3-642-07356-4 ISBN 978-3-662-09897-4 (eBook)
DOI 10.1007/978-3-662-09897-4

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag Berlin Heidelberg GmbH.

Violations are liable to prosecution under the German Copyright Law.

springeronline.com

© Springer-Verlag Berlin Heidelberg 2004

Originally published by Springer-Verlag Berlin Heidelberg New York in 2004

Softcover reprint of the hardcover 1st edition 2004

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Data conversion: PTP-Berlin Protago-TeX-Production GmbH, Berlin

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper 57/3141/ts 5 4 3 2 1 0

Preface

Silicon. The evolution and development of humanity are commonly characterized by the key words Stone Age, Bronze Age, and Iron Age; that is, characterized by materials. Curse or benefit to mankind? The discovery and utilization of semiconductors, particularly of silicon, revolutionized our living conditions, society, social life, and maxims in a few years, even more than what happened during all the material-specified periods before. Perhaps, one day, our descendants will call the period at whose beginning we live the Silicon Age. However, to be correct, the present period is characterized of the discovery and development of a whole bunch of new materials and their utilization. These materials are new alloys, ceramics, the plastics and synthetics produced by organic chemistry, composites, biomaterials, and the materials of microelectronics, nanotechnology, and space science. The materials of microelectronics are silicon, other elemental semiconductors, compound semiconductors, and organic semiconductors. With regard to the interdependences of these materials and their utilization, silicon plays a central role as one of the base materials for electronics. Have we lived in the Silicon Age for only half a century and already jumped into a new age of synthetic organic materials for electronics? We do not know.

The first intensive work on silicon started more than 50 years ago. One of the European semiconductor laboratories was installed by the industry in a centuries-old, little countryside castle in Pretzfeld, in the north-east of Bavaria, Germany. Clean-room conditions corresponding to today's definition were unknown. However, a small but brilliant team of scientists and technicians working with enthusiasm in this place developed techniques to produce some of the most outstanding silicon at that time. This laboratory in the heart of Europe became a key laboratory for single-crystal silicon. It was an impressive and exciting occasion 25 years later to visit as a younger scientist (EFK) an older scientist (Spenke), one of the leading pioneers of the first hours concerning silicon, in the castle at Pretzfeld and to be informed by him about the battles of the past. *It was not trial and error, it was trial and error and trial and success. It was more the spirit than the money.* That is the tenor of the writing of this little handbook, the background to presenting a report on silicon starting from the infancy of its use and finishing with developments in the newest exciting fields. It means learning from the

past but also looking with an open mind beyond the rim of the dish into the future. There is not only silicon!

Colleagues who themselves worked or are working in research, development, and technology in the relevant fields were asked, worldwide, to contribute as authors and coauthors. The selection of the topics had necessarily to be limited in order to obtain an easy-to-survey content. We accept that other editors would perhaps have made a different selection. We shall live with that!

Finally, we would like to thank the authors for writing their contributions and for their helpful and understanding cooperation during the editing of this book.

Particular thanks are due to W. Freiesleben for establishing contacts with potential contributors.

Strasbourg, Munich
October 2003

*Paul Siffert
Eberhard F. Krimmel*

Contents

1 Introduction: Silicon in All Its Forms

<i>J. Chelikowsky</i>	1
1.1 Introduction	1
1.2 From the 1960s to the Early 1970s: Energy Bands	2
1.3 1970s: Theory of Surfaces Applied to Silicon	8
1.4 1980s: Structural Energy of Silicon	11
1.5 1990s: Structure and Electronic Properties of Silicon Clusters and Quantum Dots	16
1.6 The Future	20

Part I “The” Crystalline Bulk Semiconductor Silicon

2 Silicon: the Semiconductor Material

<i>W. Heywang, K.H. Zaininger</i>	25
2.1 Introduction	25
2.2 Early History	26
2.3 Competition and Cooperation in the Silicon Race	28
2.4 Initial Device Applications	34
2.5 MOS Technology and Integration	36
2.6 Conclusion	39
2.6.1 Silicon	40
2.6.2 Silicon Dioxide	41
2.6.3 Si–SiO ₂ Interface	41

3 Silicon: an Industrial Adventure

<i>E.F. Krimmel</i>	43
3.1 Introduction	43
3.2 The Principal Processes	43
3.3 Some Details of Silicon Production	44

Part II Polycrystalline Silicon

4 Polycrystalline Silicon Films for Electronic Devices

<i>A. Slaoui, P. Siffert</i>	49
4.1 Introduction	49
4.2 Classification of Polycrystalline Silicon Films	50
4.3 Growth and Microcrystalline Structure of Poly-Si	51
4.3.1 Poly-Si by CVD	52
4.3.2 Poly-Si by Crystallization of Amorphous Si	59
4.3.3 Chemistry of Grain Boundaries in CVD Poly-Si	62
4.3.4 Doping of Poly-Si	63
4.4 Electronic Properties of Poly-Si	64
4.5 Conclusion	68

5 Silicon for Photovoltaics

<i>J.-C. Muller, P. Siffert</i>	73
5.1 Introduction	73
5.2 Silicon Material for PVs	74
5.2.1 History and State of the Art of Different Growing Processes	75
5.2.2 The Thin-Film Deposition Processes	79
5.3 Transport Properties in PV Silicon	82
5.3.1 Effects of Defects and Impurities on the Transport Properties of Silicon	83
5.3.2 Improvement of the Material by Gettering	84
5.4 Silicon Solar Cells	87
5.4.1 Silicon Solar Cell Technology in Comparison with Other Technologies	87
5.4.2 Multicrystalline Silicon Solar Cell Technology	88
5.5 Conclusion	89

Part III Epitaxy, Films, and Porous Layers

6 Films by Molecular-Beam Epitaxy

<i>I. Eisele, J. Schulze, E. Kasper</i>	95
6.1 Equipment Principles and Growth Mechanisms	95
6.2 Historical Development	98
6.3 Stability of Strained Heterostructures	99
6.3.1 Critical Thickness of Strained Layers	100
6.3.2 Metastable Pseudomorphic Growth	101
6.3.3 Processing and Annealing of Device Structures	102
6.4 Dopant Distribution in Films Grown by Silicon MBE	102
6.4.1 The Doping Problem	102

6.4.2	Abrupt and δ -Type Doping Profiles	105
6.5	Semiconductor Device Research	106
6.5.1	Heterojunction Bipolar Transistors	107
6.5.2	SiGe MOSFETs and MODFETs	108
6.5.3	Vertical MOSFET Structures	110
6.6	Selected Research Highlights	112
6.6.1	Cascade Laser	112
6.6.2	Resurf Structures	113
6.6.3	Self-Organization and Ordering	114
6.7	Conclusion	119

7 Amorphous Hydrogenated Silicon, a-Si:H

<i>W. Fuhs</i>	123	
7.1	Introduction	123
7.2	Preparation and Structural Properties of Amorphous Silicon	125
7.3	Electronic Properties of Hydrogenated Amorphous Silicon, a-Si:H	126
7.4	Photoluminescence and Photoconductivity	130
7.5	Metastable States	132
7.6	Amorphous-Silicon Solar Cells	133

8 Silicon-on-Insulator and Porous Silicon

<i>J.-P. Colinge</i>	139	
8.1	What is Silicon-on-Insulator?	139
8.1.1	General Properties of SOI MOS Transistors	139
8.1.2	SOI Applications	140
8.2	SOI Materials	141
8.2.1	Early SOI Materials	141
8.2.2	Silicon-on-Sapphire (SOS)	142
8.2.3	SIMOX	143
8.2.4	Wafer Bonding and Etch-Back (BESOI)	150
8.2.5	Smart-Cut®	153
8.2.6	Eltran®	159
8.3	Conclusion	162

Part IV Lattice Defects

9 Defect Spectroscopy

<i>H.G. Grimmeiss</i>	171	
9.1	Introduction	171
9.2	Fundamental Parameters Characterising the Properties of Defects	173
9.3	Junction Space Charge Techniques	175
9.3.1	Capacitance Techniques	175

9.4	Optical Measurement Methods	
	Other than Junction Space Charge Techniques	182
9.4.1	Photothermal Ionisation Spectroscopy	182
9.4.2	Fourier Photoadmittance Spectroscopy	188

10 Silicon and Its Vital Role in The Evolution of Scanning Probe Microscopy

<i>F.J. Giessibl</i>	191	
10.1	Introduction	191
10.2	Silicon as a Benchmark for Scanning Probe Microscopes	191
10.3	Silicon as a Material for AFM Cantilevers	198
10.4	The Si(111) (7×7) Surface as a Probe for STM and AFM Tips	198

Part V Doping Silicon

11 Defects, Diffusion, Ion Implantation, Recrystallization, and Dielectrics

<i>E.F. Krimmel</i>	207	
11.1	Introduction	207
11.2	High-Temperature Doping by Diffusion	208
11.3	Defects and Diffusion Mechanisms	209
11.3.1	Lattice Defects, Diffusion, and Gettering	210
11.4	Ion Implantation	211
11.5	The Special Case of Silicon, Nitrogen, Carbon, and Dielectrics	214
11.6	Implantation Profiles	216
11.7	Sputtering and Profiles	218
11.8	Radiation-Induced Defects and States, Surface States, and Interface States	219
11.9	Annealing of Ion-Implanted Specimens	220
11.9.1	Furnace Annealing	220
11.9.2	Electron Beam, Laser Beam, and Rapid Thermal Annealing	222
11.10	Conclusion	226

12 Neutron Transmutation Doping (NTD) of Silicon

<i>M. Schnöller</i>	231	
12.1	Introduction	231
12.2	Conventional Phosphorus Doping	231
12.3	Phosphorus Doping by Means of Neutron Irradiation	233
12.3.1	History	233
12.3.2	Doping Reactions	234
12.3.3	Side Reactions	235

12.3.4	Radioactivity of the Irradiated Silicon	236
12.3.5	Annealing of Crystal Defects	237
12.3.6	Technological Implementation of Silicon Doping	237
12.4	A Forward Look	240

Part VI The Roles of Certain Impurities

13 Transition Metal Impurities in Silicon

<i>T. Heiser</i>	245	
13.1	Introduction	245
13.2	Diffusion and Solubility	246
13.3	Electrical Activity	251
13.4	Impurity Engineering	254
13.4.1	Gettering	254
13.4.2	Trace Detection	256
13.4.3	Other Engineering Issues	257
13.5	Conclusion	258

14 Hydrogen

<i>C.A.J. Ammerlaan</i>	261	
14.1	Introduction	261
14.2	Hydrogen Atoms and Molecules	263
14.3	Passivation of Acceptors	269
14.4	Passivation of Donors	273
14.5	Transition-Metal–Hydrogen Complexes	277
14.6	Conclusion	281

Part VII Devices

15 Power Semiconductor Devices

<i>A. Porst</i>	293	
15.1	Introduction	293
15.1.1	History	293
15.1.2	Requirements on Power Semiconductor Devices	296
15.2	Diode	298
15.2.1	Blocking-Voltage Capability (Reverse or Blocking State)	298
15.2.2	Conducting (Forward) State	302
15.2.3	Dynamic Behavior	305
15.2.4	Trends	308
15.3	Thyristor	309
15.3.1	Behavior in Principle	309
15.3.2	Blocking-Voltage Capability	310
15.3.3	Conducting State	311

15.3.4	Dynamic Behavior	312
15.3.5	Trends	314
15.4	GTO (Gate Turn-Off Thyristor)	315
15.4.1	Behavior in Principle	315
15.4.2	Dynamic Behavior	315
15.4.3	Trends	317
15.5	Bipolar Transistor	317
15.5.1	Behavior in Principle	317
15.5.2	Conducting State	318
15.5.3	Blocking-Voltage Capability	319
15.5.4	Dynamic Behavior	319
15.5.5	Trends	320
15.6	MOS Transistor (Metal–Oxide–Silicon Transistor)	321
15.6.1	Behavior in Principle	321
15.6.2	Conducting State	323
15.6.3	Dynamic Behavior	325
15.6.4	Trends	328
15.7	IGBT (Insulated-Gate Bipolar Transistor)	329
15.7.1	Behavior in Principle	329
15.7.2	Conducting State	331
15.7.3	Blocking-Voltage Capability	333
15.7.4	Dynamic Behavior	333
15.7.5	Trends	335
15.8	Conclusions	336
16 Compensation Devices Break the Limit Line of Silicon		
<i>G. Deboy</i>	341
16.1	Introduction	341
16.2	Today's High-Voltage Device Concepts and the Way Towards the Compensation Principle	342
16.3	Manufacturing Technology and Its Challenges	347
16.4	Characteristics of Compensation Devices	350
16.5	What Is the Impact on Typical Power MOSFET Applications?	358
16.6	Conclusion and Outlook	360
17 Integrated Circuits		
<i>J. Borel</i>	363
17.1	Introduction	363
17.2	A Jump into the Past	363
17.3	Importance Gained by ICs in the Global Economy	364
17.4	The Market Constraints	366
17.5	The Products “Enablers”	367
17.6	Integration Capabilities	368
17.7	Design Bottlenecks	370
17.8	Application Domains and Product Families	372

17.8.1	Application Domains	372
17.8.2	Changes in Product Families	372
17.9	Conclusion	373
18 Silicon Nanoelectronics: the Next 20 Years		
<i>L. Risch</i>		375
18.1	Introduction	375
18.2	CMOS Scaling	375
18.3	Novel MOSFETs Below 50 nm	377
18.3.1	Strained SiGe	377
18.3.2	Strained Silicon	379
18.3.3	Vertical Transistors	381
18.3.4	Partially and Fully Depleted SOI	383
18.3.5	Double-Gate Transistors	386
18.4	FinFET Memory Cell	388
18.5	Limits of Si MOSFETs	390
18.6	Emerging Devices	392
18.6.1	Single-Electron Transistors	392
18.6.2	Molecular Devices	394
18.6.3	Carbon Nanotubes	394
18.7	Perspectives	395
19 Lithography for Silicon Nanotechnology		
<i>J. Kretz</i>		399
19.1	Introduction	399
19.2	Optical Lithography	400
19.3	Next-Generation Lithographies	403
19.4	Electron Beam Lithography	405
19.5	Nanoimprint Methods	410
19.6	Proximal-Probe Lithography	411
19.7	Conclusion	413
20 Silicon Sensors		
<i>E.F. Krimmel</i>		415
20.1	Introduction	415
20.2	“Chemical” Sensors	415
20.2.1	Biosensors	418
20.3	“Physical” Sensors	418
20.4	New Ideas and Developments	419

Part VIII Supplementing Silicon: the Compound Semiconductors

21 Supplementing Silicon: the Compound Semiconductors

<i>M. Jurisch, H. Jacob, T. Flade</i>	423
21.1 Introduction	423
21.2 The Hard Way to the Successful Product	424
21.3 Properties of III–V Compounds	427
21.4 III–V-Based Devices, Device Technologies, and Requirements for Substrates	429
21.5 GaAs: from Synthesis of the Compound to Wafers	432
21.5.1 Basic Considerations	432
21.5.2 GaAs Synthesis	439
21.5.3 Crystal Growth	441
21.5.4 Heat Treatment	446
21.5.5 Assessment of Crystals	447
21.5.6 Wafering	453
21.6 The Future of GaAs and III–V Compounds	455

Part IX New, Exciting Fields: Do They Amalgamate with Silicon?

**22 Quantum Computation by Electron Spin
in SiGe Heterostructures**

<i>F.A. Baron, K.L. Wang</i>	465
22.1 Introduction	465
22.2 The Expected Performance and Device Physics Issues of the QC	467
22.3 SiGe Implementation of the QC Using Electron Spin	468
22.4 Alternative Proposals	473
22.4.1 Pure Si Quantum Dots for QC	473
22.4.2 GaAs and CdTe Quantum Dots for QC	474
22.5 Conclusion	475

23 Carbon Nanotube Applications in Microelectronics

<i>W. Hoenlein, F. Kreupl, G.S. Duesberg, A.P. Graham, M. Liebau, R. Seidel, E. Unger</i>	477
23.1 Introduction	477
23.2 Nanotube Fabrication	479
23.3 Carbon Nanotube Interconnects	480
23.4 Carbon Nanotube Transistors and Circuits	482
23.5 CNTFET Simulations and Vertical-CNTFET Concept	483
23.6 Conclusions	487

24 Creating Systems for Ambient Intelligence

<i>K. Delaney, J. Barton, S. Bellis, B. Majeed, T. Healy, C. O'Mathuna, G. Crean</i>	489
24.1 Introduction	489
24.2 The Role of Silicon	490
24.2.1 Modular Computational Platforms	492
24.2.2 Microelectromechanical Systems	492
24.2.3 New Silicon Form Factors	493
24.3 Ambient Intelligence: Developmental Methodology	494
24.4 Novel Computational Solutions and Systems	495
24.4.1 The Disappearing Computer	495
24.4.2 Extrovert Gadgets	496
24.5 Novel Advanced Integration Technologies	496
24.5.1 Intelligent Seed Programme	497
24.6 New Silicon Forms for Sensors and Actuators	503
24.6.1 Fibre Computing Technology	503
24.7 Conclusions	510

25 Semiconductors with Brain

<i>P. Fromherz</i>	515
25.1 Introduction	515
25.2 Iono-electronic Interfacing	515
25.2.1 Planar Core-Coat Conductor	517
25.2.2 Cleft of Cell-Silicon Junction	517
25.2.3 Conductance of the Cleft	519
25.2.4 Ion Channels in Cell-Silicon Junction	521
25.3 Neuron-Silicon Circuits	521
25.3.1 Transistor Recording of Neuronal Activity	522
25.3.2 Capacitive Stimulation of Neuronal Activity	524
25.3.3 Circuits with Two Neurons on Silicon Chip	525
25.4 Brain-Silicon Chips	528
25.4.1 Tissue-Sheet Conductor	528
25.4.2 Transistor Recording of Brain Slice	529
25.5 Summary and Outlook	531
Index	533

List of Contributors

Ammerlaan, C.A.J

Van der Waals-Zeeman Institute
University of Amsterdam
Valckenierstraat 65
1018 XE Amsterdam,
The Netherlands
ammerlaan@science.uva.nl

Baron, F.A.

University of California Los Angeles
CA 90095-1594, USA
baron@ee.ucla.edu

Barton, J.

NMRC Institute, Lee Maltings
Prospect Row
Cork, Ireland
john.barton@nmrc.ie

Bellis, S.

NMRC Institute, Lee Maltings
Prospect Row
Cork, Ireland
stephen.bellis@nmrc.ie

Borel, J.

JB-R&D
12 rue du Drac
38120 Saint Egreve, France
josephborel@aol.com

Chelikowsky, J.

Department of Chemical Engineering
and Materials Science
University of Minnesota
Minneapolis, MN 55455, USA
jrc@msi.umn.edu

Colinge, J.-P.

Department of Electrical and
Computer Engineering
University of California
Davis, CA 95616, USA
colinge@ece.ucdavis.edu

Crean, G.

NMRC Institute, Lee Maltings
Prospect Row
Cork, Ireland
gabriel.crean@nmrc.ie

Deboy, G.

Infineon Technologies AG
Automotive and Industrial Division
Power Management & Supplies
Balanstr. 59
81541 München, Germany
gerald.deboy@infineon.com

XVIII List of Contributors

Delaney, K.

NMRC Institute, Lee Maltings
Prospect Row
Cork, Ireland
kieran.delaney@nmrc.ie

Duesberg, G.S.

Infineon Technologies AG
Corporate Research, Nano Processes
Otto-Hahn-Ring 6
81739 Munich, Germany

Eisele, I.

Universität der Bundeswehr
München
Institut für Physik
85579 Neubiberg, Germany
ignaz.eisele@unibw-muenchen.de

Flade, T.

Freiberger Compound
Materials GmbH,
Am Junger Löwe Schacht 5
09599 Freiberg, Germany
info@fcml-germany.com

Fromherz, P.

Max-Planck-Institut für Biochemie,
Am Klopferspitz 18 A
82152 Martinsried, Germany
fromherz@biochem.mpg.de

Fuhs, W.

Hahn Meitner Institut
Abteilung Silizium – Photovoltaik
Kekulestraße 5
12489 Berlin, Germany
fuhs@hmi.de

Giessibl, F.J.

Experimentalphysik VI, EKM
Institute of Physics
University of Augsburg
86135 Augsburg, Germany
franz.giessibl@physik.uni-augsburg.de

Graham, A.P.

Infineon Technologies AG
Corporate Research, Nano Processes
Otto-Hahn-Ring 6
81739 Munich, Germany

Grimmeiss, H.G.

Solid State Physics
University of Lund
221 00 Lund, Sweden
hermann.grimmeiss@ftf.lth.se

Healy, T.

NMRC Institute, Lee Maltings
Prospect Row
Cork, Ireland
thomas.healy@nmrc.ie

Heiser, T.

Laboratoire de Physique et
Applications des Semiconducteurs
BP20
67037 Strasbourg Cedex 2, France
thomas.heiser@phase.c-strasbourg.fr

Heywang, W.

Siemens Corporate Research
Laboratories
Munich, Germany, retired
Schwabener Weg 9 A
85630 Neuherberg, Germany
walter.heyawang@t-online.de

Hoenlein, W.

Infineon Technologies AG
Corporate Research, Nano Processes
Otto-Hahn-Ring 6
81739 Munich, Germany
wolfgang.hoenlein@infineon.com

Jacob, H.

Chemitronik,
Marienbergerstrasse 17
84489 Burghausen, Germany,
herbert.g.jacob@t-online.de

Jurisch, M.

Freiberger Compound
Materials GmbH,
Am Junger Löwe Schacht 5
09599 Freiberg, Germany
jurisch@fcm-germany.com

Kasper, E.

Institut für Halbleitertechnik
Universität Stuttgart
Pfaffenwaldring 47
70569 Stuttgart, Germany
icsi2@iht.uni-stuttgart.de

Kretz, J.

Infineon Technologies AG
Corporate Research
Otto-Hahn-Ring 6
81739 Munich, Germany
johannes.kretz@infineon.com

Kreupl, F.

Infineon Technologies AG
Corporate Research, Nano Processes
Otto-Hahn-Ring 6
81739 Munich, Germany

Krimmel, E.F.

Siemens Corporate
Research Laboratories
Munich, Germany, retired
and J.W. Goethe University,
Frankfurt/Main, Germany, retired
and European Materials Research
Society
Strasbourg, France
Mendelssohnstrasse 7
82049 Pullach/Isartal, Germany
520079789984-0001@t-online.de

Liebau, M.

Infineon Technologies AG
Corporate Research, Nano Processes
Otto-Hahn-Ring 6
81739 Munich, Germany

Majeed, B.

NMRC Institute, Lee Maltings
Prospect Row
Cork, Ireland
bivragh.majeed@nmrc.ie

Muller, J.-C.

CNRS, Laboratoire PHASE, BP 20
67037 Strasbourg Cedex 2, France
jean-claude.muller@
phase.c-strasbourg.fr

O'Mathuna, C.

NMRC Institute, Lee Maltings
Prospect Row
Cork, Ireland
cian.omathuna@nmrc.ie

Porst, A.

Siemens Corporate
Munich, Germany, retired
Kurt Floericke Strasse 14
81249 München, Germany

Risch, L.

Infineon Technologies AG
Corporate Research,
Otto-Hahn-Ring 6
81739 Munich, Germany
lothar.risch@infineon.com

Schnöller, M.

Siemens Corporate
Munich, Germany
retired,
Grete-Hoffmann-Weg 24
85778 Haimhausen, Germany
m.schnoeller@web.de

Schulze, J.

Universität der Bundeswehr
München
Institut für Physik
85579 Neubiberg, Germany

Seidel, R.

Infineon Technologies AG
Corporate Research, Nano Processes
Otto-Hahn-Ring 6
81739 Munich, Germany

Siffert, P.

Laboratoire PHASE-CNRS and
European Materials Research
Society
23 rue du Loess
67037 Strasbourg Cedex 2, France
emrs@phase.c-strasbourg.fr

Slaoui, A.

Laboratoire PHASE-CNRS
23 rue du Loess
67037 Strasbourg Cedex 2, France
slaoui@phase.c-strasbourg.fr

Unger, E.

Infineon Technologies AG
Corporate Research, Nano Processes
Otto-Hahn-Ring 6
81739 Munich, Germany

Wang, K.L.

University of California Los Angeles
CA 90095-1594, USA and
Hong Kong University of Science
and Technology
Clear Water Bay
Hong Kong
wang@ee.ucla.edu

Zaininger, K.H.

Siemens Corporate Research
Laboratories
Princeton, NJ, USA, retired
9 East Shore Drive,
Princeton-Atlantic City,
Princeton, NJ 08540, USA
karl@zaininger.com

1 Introduction: Silicon in All Its Forms*

J. Chelikowsky

1.1 Introduction

I am truly and deeply honored to be selected to give this talk. One of the highlights of my early career was to take a trek up to Harvard and visit with Professor David Turnbull. Although I only spent a small amount of time with him, it made a considerable impact on me.

With respect to the matter at hand, I am going to tell you about my favorite material: element 14, silicon. As you know, we live in the age of silicon; it is all around us in terms of electronic devices. Silicon is the quintessential electronic material. It is often said that understanding silicon and its role in electronic materials is similar to understanding iron and its role in metallurgy and making steel.

It has not always been that way, though. On May 10, 1954, a press release was issued on the first silicon transistor. Texas Instruments announced a “revolutionary new electronic product – long-predicted and awaited: the silicon transistor.” By using silicon instead of germanium, the initial commercial silicon transistor immediately raised power outputs and doubled the operating temperature, [1].

In 1954, we had a transistor made of silicon for the first time. In 1965, we had a chip containing ~ 2000 transistors. In 2001, the Pentium 4 processor made by Intel consisted of 42 million transistors. Intel believes that by the year 2007, it will have created a processor containing one billion transistors, [2]. This is an amazing progression of technology, and we could spend days talking about the technical developments of silicon and devices made with silicon. However, I am going to discuss a different aspect of silicon technology, namely, the role that silicon has played in the development of theoretical tools for understanding materials.

In the 1970s, approximately 30,000 papers were published with the word “silicon” in the abstract. In the 1980s, the number was up to 84,000 papers. By the time we got into the 1990s, the corresponding number was 135,652 (an average of one paper every 90 minutes), [3]. This amounts to a database of roughly a quarter of a million papers over the last 30 years directly or indirectly written about silicon. Most of these papers are focused on technological

* Reproduced by permission of MRS Bulletin. J. Chelikowsky, “Silicon in All Its Forms”, MRS Bulletin, Vol. 27 No. 12 (2002), pp. 951–960.

and experimental aspects of silicon, but this body of work has profoundly influenced the theoretical community. Specifically, theorists have capitalized on this vast database. Any new theory related to electronic materials is almost always first tested and assessed against the silicon database.

I am going to cover about 40 years' worth of history in an overview of the role that silicon has played in our understanding of electronic materials. Needless to say, I will be brief in my discussion and it will be highly abridged. I will outline some major achievements over the last few decades in our fundamental understanding of electronic materials. For example, in the early 1960s, we did not know the detailed energy bands of silicon. However, by the 1970s, we knew the energy-band structure for crystalline silicon, and we moved on to more complex issues. For example, we began looking at the surfaces of silicon. In the 1980s, we addressed issues associated with defects, and by the 1990s, we were looking at liquids, clusters, amorphous solids, and other complex systems. Today, we can add nanoparticles of silicon to this list.

Before we begin this brief overview, I want to reemphasize how important silicon has been in this progression of understanding. Suppose that the technological development of the silicon transistor in 1954 had failed; would we have “Germanium Valley” today? Would we have “Germanium Graphics Inc.” instead of “Silicon Graphics Inc.”? One thing I know: we would not have advanced as much as we have in our understanding of matter without the concurrent advances in silicon technology.

1.2 From the 1960s to the Early 1970s: Energy Bands

One of the early issues in silicon research was understanding the spatial and energetic distribution of electrons within a semiconductor. This entails solving the so-called energy-band problem, that is, determining the energy of an electron as a function of its wave vector \mathbf{k} . The wave vector is a quantum number for tagging the quantum states of an electron in a crystal, just as one has the quantum numbers nlm to label a quantum state in an atom.

We have all the necessary equations and knowledge of the interactions between particles to arrive at a quantitative understanding of the electronic states of an electron in a solid. Specifically, we can write down the full many-body, time-dependent Schrödinger equation:

$$\begin{aligned} H(r_1, r_2, \dots; R_1, R_2, \dots; t) \Psi(r_1, r_2, \dots; R_1, R_2, \dots; t) \\ = i\hbar \frac{\partial}{\partial t} \Psi(r_1, r_2, \dots; R_1, R_2, \dots; t), \end{aligned} \tag{1.1}$$

where H is the full all-electron Hamiltonian representing all of the electronic interactions between electrons and nuclei, r is the coordinate of the electron, R is the coordinate of the nucleus, t is time, and Ψ is the wave function, which contains information on the spatial distribution of each electron in

the system. A solution of this equation would yield all of the information we need to know. This was recognized in 1929 by the famous theoretical physicist P.A.M. Dirac. He said that in principle, we have all of the underlying physical laws necessary for the mathematical understanding of a large part of the physics and the whole of the chemistry as contained in (1.1), but that this equation is too complex to solve. He advised the scientific community to develop “approximate, practical methods.”, [4]. Of course, this advice leads to several questions, such as what practical approximations could we develop to help us understand the electronic and structural properties of matter? Furthermore, how would we know whether those practical approximations worked? This is a key point of my presentation: *we can test any proposed approximations against the database of papers on silicon.*

Some practical approximations come immediately to mind. For example, we can remove the nuclei from consideration, and focus only on the electronic degrees of freedom. This is the so-called Born–Oppenheimer approximation, [5]. It is justified because the nuclei are very massive, compared with the electrons, and as such, the electrons can be thought to respond instantaneously to the nuclear positions. Another approximation is the one-electron approximation. Within this approximation, each electron is thought to move in the average potential field of all of the other electrons in the system. This is often called the Hartree approximation, [6]. Within these approximations, we need to solve the following equation:

$$\left[\frac{-\hbar^2 \nabla^2}{2m} + V(r) \right] \Psi_n(r) = E_n \Psi_n(r), \quad (1.2)$$

where $V(r)$ represents the electrostatic potential, m is the mass of the electron, and n is a quantum number. At this time, we will neglect any external fields and assume that the potential is not a function of time. The eigenvalues (E_n) of this equation will give us the energetic distribution of electrons, and the eigenfunctions (Ψ_n) will provide the spatial distribution of the electrons.

We have transformed the original description of the electronic structure problem (1.1) to a much simpler one. Specifically, we have replaced the many-body problem in which the wave functions involve N coordinates (for N electrons), by one in which we solve for N one-electron orbitals.

To confirm the success of these approximations, we can put the silicon test to work. But first, we need to make one additional – and a key – approximation: the pseudopotential approximation, [7–10]. We can think of matter as consisting of the nucleus, core electrons, and valence (outer) electrons. This is illustrated in Fig. 1.1. Only the outer electrons ($3s^2 3p^2$) need to be considered for our purposes because they are the chemically active electrons. Within the pseudopotential approximation, the core electrons are removed from the problem by transcribing the “all-electron” potential to a pseudopotential.

The physical content of the pseudopotential is the same as the periodic table. The periodic table is organized by columns based on the valence electrons (e.g., carbon, silicon, germanium, tin, and lead are grouped together). The

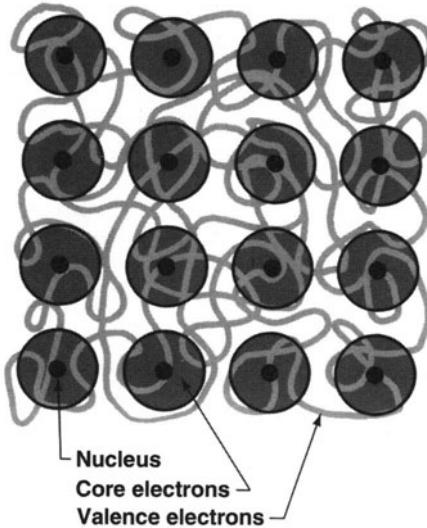


Fig. 1.1. Pseudopotential model of a solid. The nucleus and core electrons make a chemical inert ion core, while the valence (outer) electrons are the chemically active electrons

pseudopotential approximation proceeds in a similar fashion. Within this approximation, the numerical complexities of silicon and lead are comparable, even though in terms of the total number of electrons, silicon has 14 electrons and lead has 82 electrons.

In the 1960s, pseudopotentials were first used to describe the electronic structure of silicon. Since the pseudopotential replaces the *strong* all-electron potential with a *weak* pseudopotential, we can use sines and cosines to solve the problem, that is, we can use a Fourier series for both the pseudopotential and the wave functions. Because the potentials and wave functions are smoothly varying with no singularities, a Fourier transform converges quickly, [11]. This is illustrated schematically in Fig. 1.2, where a real-space pseudopotential is illustrated, along with its Fourier transform.

For silicon, we can fix the required Fourier components of the potential by fitting them to experimental data. Typically, this can be done with three parameters, and two of them are not really linearly independent. Once the potential is determined, we can extract the energy bands of silicon. The method by which one fits the pseudopotential is called the empirical pseudopotential method (EPM), [11, 12].

Silicon occurs in the diamond structure, as shown in Fig. 1.3. Within the primitive cell, there are eight electrons, which means there are four occupied valence bands. Shown in Fig. 1.4 are the energy bands for silicon in the diamond crystal structure. This represents one of the first realistic energy-band structures, [13]. In silicon, we find a bandgap of about 1.1 eV and a valence-band width of about 12 eV, [11]. I should add that this is a very easy

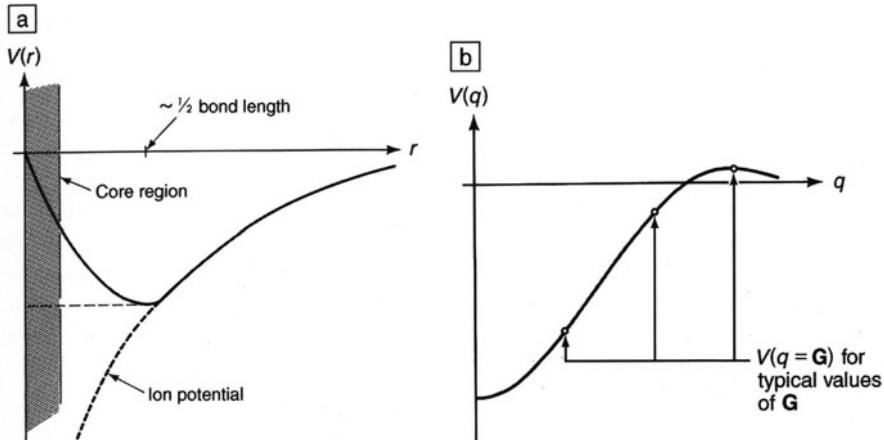


Fig. 1.2. Schematic pseudopotential in **a** real space and **b** Fourier space. $V(r)$ is the electronic potential, q is the reciprocal distance, and \mathbf{G} is a reciprocal space vector

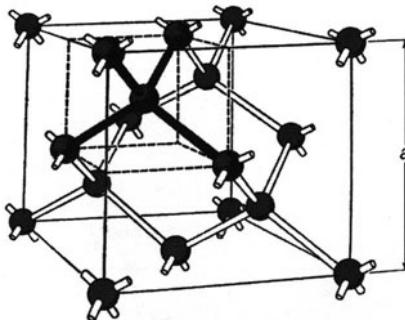


Fig. 1.3. Schematic illustration of silicon in the diamond structure

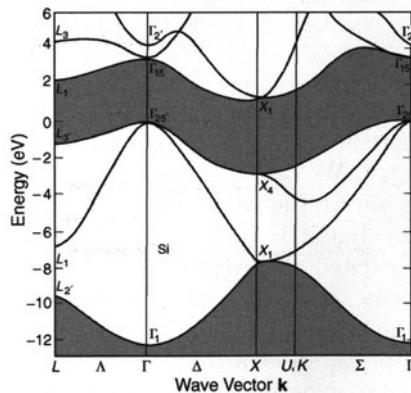


Fig. 1.4. Energy bands for silicon in the diamond crystal structure, [13]. The top of the valence band is taken as the zero of energy

numerical calculation. For example, one could do this calculation on a laptop or pocket PC; one only has to solve a 50×50 matrix problem.

As a result of the EPM calculation, we have the energetic distribution of the electrons within the crystal through the band structure, and we have the corresponding wave functions for the spatial distribution. In Fig. 1.5, I have reproduced the calculated and measured modulated reflectivity spectra for crystalline silicon, [14]. We can use the oscillator strength between filled states and empty states, along with the topology of the energy bands, to predict the origin of any structure in the reflectivity spectrum. The experimental work by Zucca and Shen in the 1970s (Fig. 1.5) reveals some highly structured features, [14]. The general agreement between experiment and theory confirms our practical approximations.

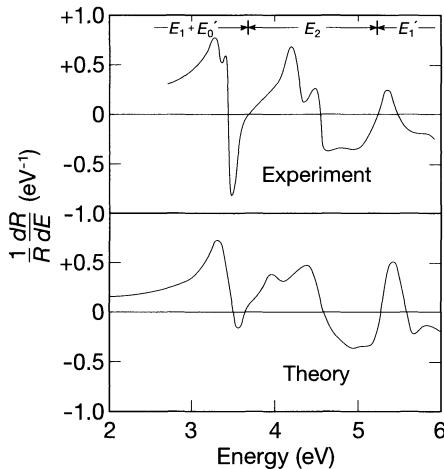


Fig. 1.5. Calculated (theory) and measured (experiment) modulated reflectivity spectra for crystalline silicon. The theoretical work is from an empirical pseudopotential calculations, [13]. The measured reflectivity is from Zucca and Shen, [14]

However, we do not get everything right. For example, in the low-energy part of the spectra, there are some sharp structures that are not reproduced in the theoretical spectra. We can attribute this absence to the lack of any spin-orbit interactions within the EPM spectra. This feature was omitted in these early calculations, but it is easy to include it. Also, excitonic effects, the interaction of the excited electron and hole, are not properly included, which results in line shapes not in complete agreement with experiment, but this is a detail. With the advent of the EPM, it was quickly realized that one could obtain reliable optical spectra from theory. These theoretical spectra could be directly related to experiment. Before the 1960s, it was controversial whether one could learn anything from the optical properties of solids. It was thought that the one-electron picture was too crude. However, by the 1970s,

it was clear that one could learn a great deal. Thus, an explosion of spectroscopic measurements of solids using optical probes began. The combination of experiment coupled to the EPM led to the first quantitative understanding of energy bands. This development had profound effects on our ability to do bandgap engineering, that is, to design an electronic material for a specific application, [11].

In Fig. 1.6, I illustrate x-ray photoemission spectroscopy (XPS) spectra from Ley et al., circa 1974, [15]. Using XPS, one can measure directly the energy distribution of occupied states in silicon (i.e., the density of states), [11]. The theoretical density of states can include both empty and filled states; the empty states are not observed in photoemission. The agreement between theory and experiment is quite good, especially given that only a few parameters were used. As in the case of the reflectivity calculations, there were some interesting details that we could not replicate in the early days. These features, such as the shoulder around -1 eV , below the valence-band maximum in the bottom graph of Fig. 1.6, occurred because of surface states. These early calculations did not include surface contributions.

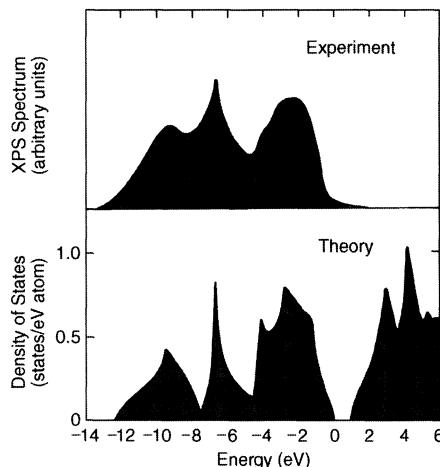


Fig. 1.6. X-ray photoemission spectroscopy (XPS) spectra and calculated density of states for crystalline silicon, from Ley et al., [15]

Perhaps one of the most interesting aspects of these early calculations concerned the spatial distribution of electrons in silicon. From the wave functions obtained from the EPM, we can extract the spatial distribution of the bonding electrons. Fig. 1.7 is a contour map for such distributions within a silicon crystal. This figure provides one of the first visual depictions of the covalent bonds in silicon.

Around 1974, Yang and Coppens did an x-ray experiment on silicon to determine the spatial distribution of electrons, [17]. With x-rays, we cannot

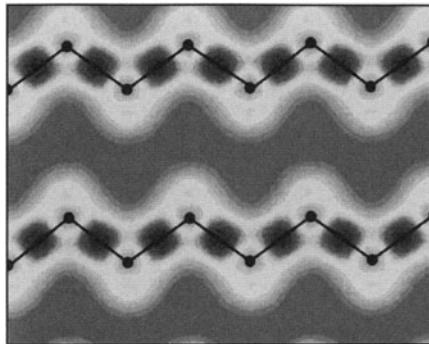


Fig. 1.7. Contour map for spatial distributions of the bonding electrons (charge density) in crystalline silicon, [11, 16]. Regions of high electron density (e.g., the covalent bonds) are shown in shades of dark gray; low-density regions are shown in pale gray going to dark gray. The plane shown is (110). The atomic positions are shown by black dots

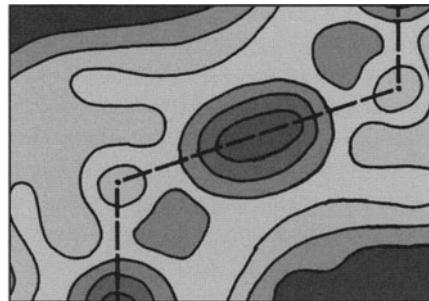


Fig. 1.8. Spatial distribution of electrons (charge density) for crystalline silicon, as inferred from x-ray measurements by Yang and Coppens, [17]. The gray coding is similar to that in Fig. 1.7

obtain the precise spatial distribution because these probes measure intensities and not amplitudes. However, with a few simple – and justifiable – approximations, Yang and Coppens extracted what they believed to be the bonding configuration, [17]. The agreement between theory and experiment was striking (see Fig. 1.8).

1.3 1970s: Theory of Surfaces Applied to Silicon

Suppose we want to understand the *surface* of silicon. Many things happen at a surface. We no longer have a periodic system perpendicular to the surface. We have threefold-coordinated Si atoms at the surface, in contrast with fourfold-coordinated atoms within the bulk crystal. We may not have the same atomic structure or the same electronic configuration at the surface

that we have in the bulk. We have to account for these differences in solving for the electronic structure.

In the mid-1970s, we introduced a new type of pseudopotential, [10, 18, 19] that resulted from the core electrons and nuclear charge and did not include any screening from the valence electrons. We fixed the positions of these *ion-core* potentials and figuratively threw in the valence electrons. The valence electrons were allowed to relax to the lowest-energy configuration as determined by using density functional theory and the so-called local density approximation, [20, 21]. This approximation allows us to screen the ion-core pseudopotential with an electrostatic potential arising from the valence electrons. This potential includes classical electrostatic screening and an exchange-correlation term that depends only on the density of a point of interest. Exchange-correlation terms arise from inherently quantum interactions, which have no classical analogue.

We solve the screening process self-consistently, that is, we generate a Hamiltonian that is consistent with the charge density, based on the wave functions. The resulting potentials are called *ab initio* pseudopotentials. To construct these potentials, we start from an atomic-structure calculation, extract the ion-core pseudopotential, and apply it to condensed-matter systems to obtain the lowest-energy solution. Unlike the empirical pseudopotentials, no input from experiment is used.

We handle the loss of periodicity created by the presence of a surface by introducing “artificial periodicity.” Imagine that I want to look at the electronic structure of a cluster – a system with no translational periodicity. I can do the following: I place the cluster in a large box and artificially repeat that box to fill up all space. When so constructed, this box is called a supercell, [17, 18, 22]. I now have a periodic system – a crystal of isolated clusters. If I solve for the electronic structure of this system, I get the electronic structure of an isolated cluster. I can use this approach for any localized system, provided that I can take a supercell of sufficient size. Moreover, I can use computer codes that have been developed over the years to look at crystals and apply them to clusters, surfaces, liquids, defects – all the forms of silicon that are important.

Let us examine one of the first such calculations for a surface. We will consider the silicon (111) ideal surface. The (111) surface is the cleavage plane for silicon. A model for this geometry is shown in Fig. 1.9. At the surface, each atom is threefold-coordinated. We can learn a few things about surfaces by considering this simple system. For example, if we examine the distributions of electronic charge at the surface (Fig. 1.10), we notice that the surface perturbation heals very quickly; the darker shades of gray regions indicating covalent bonds (which are missing at the surface) are restored (“healed”) a short distance below the surface. Within a few bond lengths, the charge density assumes the same distribution as it had in the infinite crystal. We can also imagine how impurities might migrate from the surface into the bulk crystal. For example, a channel devoid of any charge density exists from the vacuum into the bulk crystal, [22].

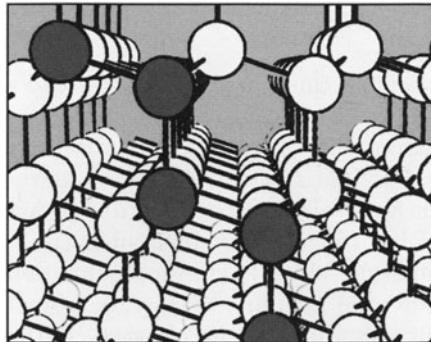


Fig. 1.9. Model of the geometry of the Si(111) ideal surface. The (111) plane corresponds to the top edge of the figure. The atoms shown as dark gray circles are within a plane shown in Fig. 1.10

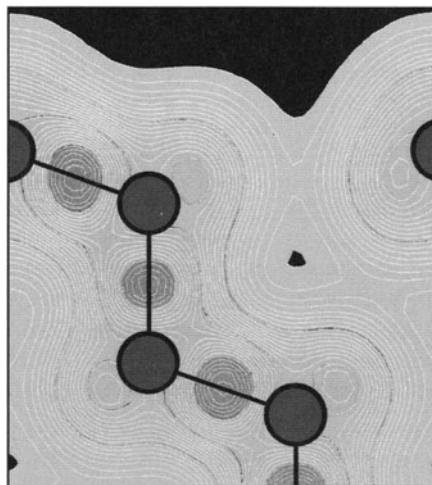


Fig. 1.10. Charge density of the Si(111) surface, [22]. The dark gray circles represent the atoms that correspond to the plane shown in Fig. 1.9. Here, regions of high charge density are shown in darker shades of gray

One new electronic feature exists at the surface, namely, electronic states that occur at the vacuum–solid interface. Called surface states, they decay both into the vacuum and into the bulk crystal. Electrons occupying such states are “trapped” at the interface between the crystal and vacuum. These surface states arise from dangling bonds that occur because unsaturated bonds are present at the surface, [22]. When adsorbate atoms interact with the silicon surface, the nature of the chemical bond between the adsorbate and the surface will be largely determined by the presence of these surface states.

We do not see the dangling-bond surface states in Fig. 1.10 because the charge density is found by summing up over all occupied states. Most of those

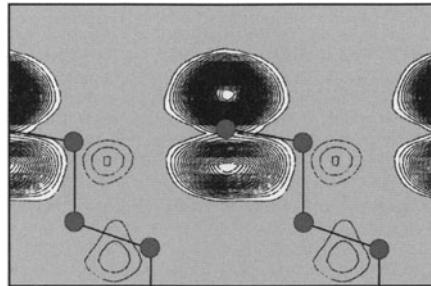


Fig. 1.11. Contour map for the dangling-bond state on the Si(111) surface. The plane is similar to that shown in Fig. 1.9. The vacuum is represented by the large gray areas. See Schlüter et al. for details, [22]

occupied states decay gently into the vacuum. However, within the bandgap of silicon, there are electronically active surface states corresponding to dangling bonds. If we examine the spatial distribution of these states, we find these states decay both into the vacuum and into the crystal, as in Fig. 1.11.

With respect to surface properties in general, we can use *ab initio* pseudopotentials to examine many issues. For example, we can look at adsorbates and the nature of the bonding. We can calculate work functions. We can see what happens when we passivate the surface. We can put metals on surfaces and look at the formation of Schottky barriers. We know this by testing all of these issues against a known silicon database, [23].

However, we have an unresolved problem. Up to this point, we have assumed a given geometry for the surface. In the case illustrated for the Si(111) surface, we assumed the atomic positions were not influenced by the presence of the surface. However, there is no good reason for making this assumption. Silicon surfaces are known to “reconstruct.” A reconstruction occurs when the surface atoms do not maintain atomic positions corresponding to the bulk symmetry of the crystal. The reconstructions for a simple elemental crystal like silicon can be quite complex. For example, the periodicity of the Si(111) reconstructed surface can be described at high temperatures by a unit cell 49 times the size of the ideal cell, that is, the so-called 7×7 reconstruction, [22]. We need to know how the atoms are arranged at the surface. How are we going to proceed? We need a method for examining energy differences between competing surface configurations.

1.4 1980s: Structural Energy of Silicon

We go back to density functional theory. This theory dates back to Fermi and Slater, but the modern version dates to seminal work by Kohn and his collaborators in the mid-1960s, [20, 21]. In this work, they found that the many-body part of the problem can be included by an effective exchange-

correlation potential. If one knows all of the appropriate electronic interactions – for example, the exchange-correlation, the ion–ion-interaction, and the electron–electron-electrostatic contributions – we can add up these interactions and find the total energy. We can calculate the energies of different structures. If we know the energies of structures, then we can predict which surface structure might be the lowest-energy one.

The key equation to be solved is from Kohn and Sham:

$$\left[\frac{-\hbar^2 \nabla^2}{2m} + V_{\text{ion}}(r) + V_{\text{H}}(r) + V_{\text{xc}}(r) \right] \Psi_n(r) = E_n \Psi_n(r), \quad (1.3)$$

where V_{ion} is the ion-core pseudopotential, V_{H} is the Coulomb or Hartree potential, and V_{xc} is the effective exchange-correlation potential. The Hartree potential is obtained by solving a Poisson equation:

$$\nabla^2 V_{\text{H}}(r) = -4\pi e \rho(r), \quad (1.4)$$

where $\rho(r)$ is the charge density given by

$$\rho(r) = -e \sum_{\text{occup}} |\Psi_n(r)|^2. \quad (1.5)$$

The summation is over all occupied states. Within the local density approximation, the V_{xc} potential is a functional of the charge density: $V_{\text{xc}} = V_{\text{xc}}[r]$. Once the equation is solved, we can find the total electronic energy of the system from

$$E_{\text{total}} = \sum_{\text{occup}} E_n - \frac{1}{2} \int V_{\text{H}} \rho d^3r + \int (E_{\text{xc}} - V_{\text{xc}}) \rho d^3r + E_{\text{ion-ion}}. \quad (1.6)$$

The first sum is over all occupied states. The second term subtracts off the double counting terms. The third term subtracts off the exchange-correlation potential and adds in the correct energy density functional. The last term is the ion–ion-core repulsion term.

In one of the first applications of this formalism to real systems, Yin and Cohen evaluated the total energies of different silicon crystals as a function of crystal volume, [24]. Their work is shown in Fig. 1.12. They considered the following possible crystal structures for silicon: diamond, hexagonal diamond, white tin (β -Sn), simple cubic, simple hexagonal, hcp, bcc, and fcc. The diamond crystal structure was the lowest in energy.

Relying on our large silicon database, we can assess whether this picture is correct. The diamond structure is the lowest-energy structure for silicon, both experimentally and theoretically. By creating a common tangent between the diamond structure curve and the white tin structure curve, one can find the transition pressure between these two forms. The slope of that tangent will give us the transition pressure. They found this to be 9 GPa, which is roughly in agreement with experiment, [24]. Perhaps more important, Cohen and his

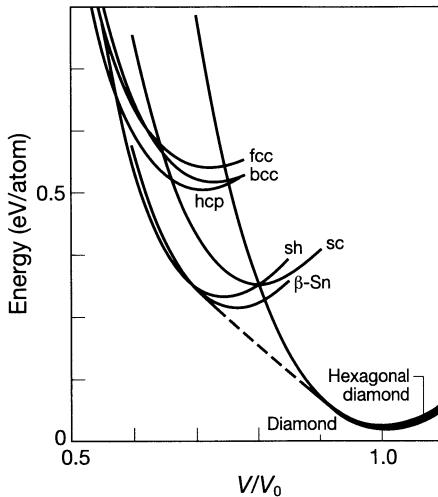


Fig. 1.12. Energy versus volume plot for several crystal structures of silicon: diamond, hexagonal diamond, white tin (β -Sn), simple cubic (sc), simple hexagonal (sh), hcp, bcc, and fcc. This work is after Yin and Cohen, [24]

students later found one of these high-pressure forms of silicon to be metallic, predicted it to be a superconductor, and it has been so observed, [25]. This stands out as one of the best examples of a theory predicting a property of silicon.

As such, we have a lot of confidence in density functional theory when we are dealing with semiconductors. Typically, we obtain lattice parameters to within 1–3%, compressibilities within 5–10%, and lattice vibrations within 1–3%.

In the first applications of total energies from density functional theory, one did not do so well with cohesive energies. A cohesive energy is the energy of an isolated silicon atom versus that same silicon atom in a crystal. Those are very different environments, when compared with structural energies. If we are simply altering the volume of a crystal by small changes in the lattice constant, and comparing one solid-state configuration to another solid-state configuration, the resulting energy changes are more reliable. Why is this? For density functional theory to work, error cancellations are important and must be largely complete. For cohesive-energy calculations, the error cancellations between dissimilar environments are not as complete as for lattice-parameter calculations. If we examine more contemporary density functional theories, such as the generalized gradient approximation, the errors in cohesive energies are considerably reduced, [26].

One problem to be addressed later is that no excited-state properties have been included in what I have shown so far. At this point, we are interested only in the ground-state properties. Ground-state properties correspond to equilibrium properties in which the electrons occupy only the lowest-energy

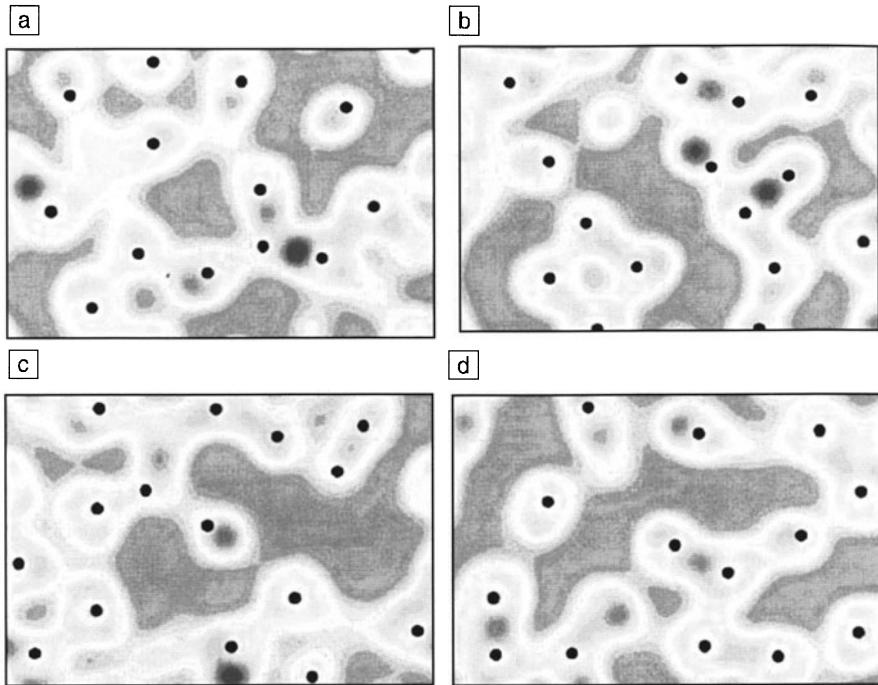


Fig. 1.13. Charge density in liquid silicon. Frames show movements of atoms in the liquid over time. The shades of pale gray areas indicate a high electron density, as in Fig. 1.7

orbitals. Such properties include the cohesive energy of a solid and the bond lengths. In contrast, if we shine light on a material, and the electrons are excited to higher-energy orbitals, we no longer have a system within the ground state. Density functional theory works only for the ground state. Later, I will deal with excited states and optical excitations.

If we know the energy as a function of structure, we can do *ab initio* molecular dynamics. By *ab initio* here, I mean that the required interatomic forces in the simulations are calculated quantum-mechanically and not fit to some experimental data. For example, the Hellman–Feynman theorem can be used to calculate the forces on a particular particle with the use of pseudopotentials and density functional theory. We can use these quantum forces to integrate the equations of motions of the nuclei without making any ad hoc assumptions about interatomic forces.

Of course, just as Yin and Cohen first looked at silicon to assess the reliability of density functional theory, the first applications of *ab initio* molecular dynamics for liquids were focused on silicon, [27, 28].

Before we examine liquids, it is worthwhile to reexamine the charge density distribution within the crystal (see Fig. 1.7). Within the plane illustrated in this figure, the charge density is highly localized in bonds. What will the

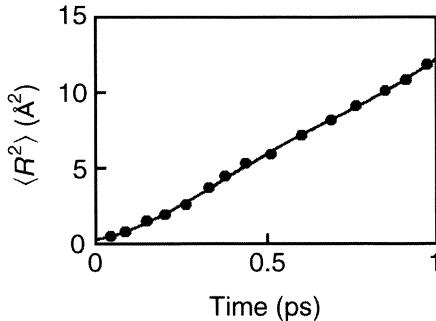


Fig. 1.14. Root-mean-square displacement of a Si atom in liquid Si. The temperature of the liquid is near the melting point. The slope of this curve is directly related to the self-diffusion constant of liquid silicon, [28]

distribution of electrons in a liquid look like? In the liquid, the atoms are moving and one can only examine “snapshots” of the system or do some sort of averaging. In Fig. 1.13, I have illustrated the charge density of the liquid state. Liquid silicon is quite different than crystalline silicon. The charge density of the liquid is quite diffuse and not highly localized. In fact, liquid silicon is observed experimentally to be metallic, as most covalent semiconductors are when melted. Yet, there is evidence for covalent bonds existing within the melt. In Fig. 1.13, you can see such a covalent bond existing between two atoms within the liquid. This finding implies that covalent bonds can exist in the silicon liquid, at least for a brief instant of time. Some have speculated that by thermal-quenching the liquid into a solid structure, we may preserve the microstructure of the melt. This would render a system with covalent bonds in a metallic matrix. Such a configuration is typical of a superconductor, not a semiconductor. However, no one has been able to cool the silicon melt fast enough to preserve the metallic structure of the liquid. So, we do not know if silicon frozen in a liquid structure would be a superconductor or not.

There are other issues of the melt that can be examined with theoretical tools. For example, one can calculate the self-diffusion in liquid silicon. We can look at the rms displacement of that silicon atom as a function of time, [28]. From the slope of this curve (see Fig. 1.14), we can obtain the self-diffusion constant. Diffusion constants are needed when modeling crystal growth, but they are difficult to measure. They are not so difficult to calculate. However, I confess this is one instance where our vast silicon database fails. I do not believe the diffusion of silicon in liquid silicon has been measured. On the other hand, it is nice to know that there are things that theory can do better than experiment.

1.5 1990s: Structure and Electronic Properties of Silicon Clusters and Quantum Dots

Clusters are a special form of matter, groups of atoms that are stable only in isolation, making them difficult to handle either experimentally or theoretically, [19]. For example, a cluster of 10 silicon atoms will not be stable if this cluster interacts with another cluster of 10 silicon atoms. The two clusters will merge seamlessly into a 20-atom silicon cluster. This would not be true if two *molecules* interacted. For example, consider two benzene molecules. If these molecules interact, they will retain their integrity as molecular species. In addition, we may have a number of atoms within the clusters interacting with no special symmetry, which means we can have a number of degrees of freedom – both structurally and electronically. This complicates the problem enormously from a theoretical perspective.

Consider the following issue: what is the structure of a silicon cluster? We cannot answer this question by examining bulk forms of silicon. The atoms in a cluster are effectively all surface-like, which means they are not fourfold-coordinated. Just as a silicon surface will reconstruct, so will a cluster of

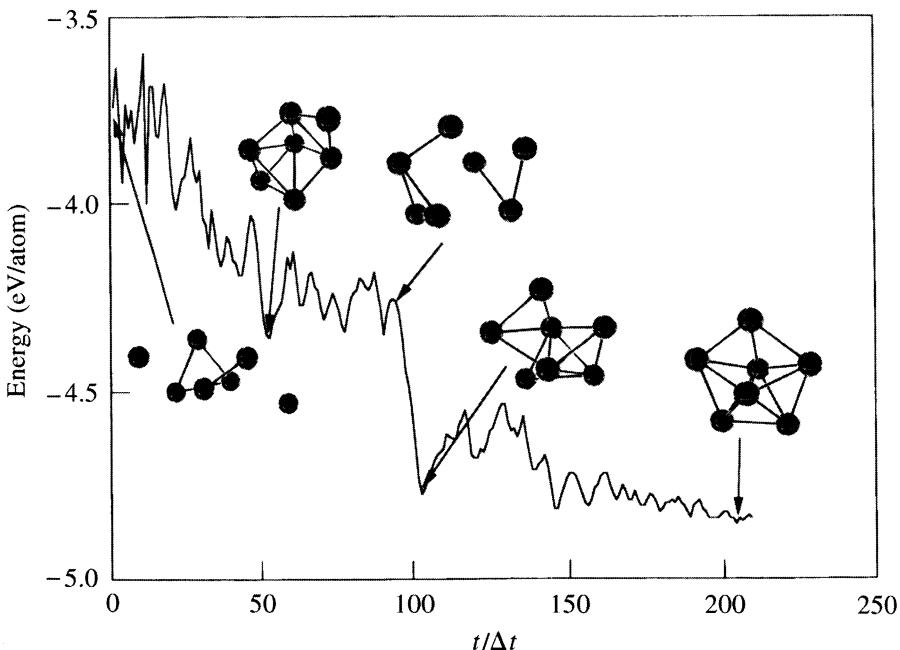


Fig. 1.15. Molecular-dynamics simulated anneal for a seven-atom silicon cluster. Over the course of this simulation, the temperature is reduced from about 5000 K (*on left*) to 300 K (*on right*). The resulting structure of the cluster agrees with other theoretical studies and experimental data. Time steps are indicated by Δt

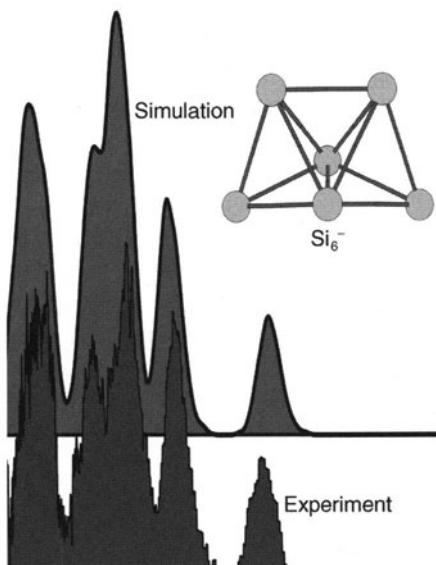


Fig. 1.16. Simulated photoemission spectrum for a six-atom, negatively charged cluster of silicon (Si_6^-), compared with experiment, [31]

silicon. We cannot measure the structure of a cluster directly. How do we know what the structure is?

We can answer this theoretically by performing a “simulated anneal.” An anneal occurs when we slowly cool a system from a high temperature to a low one. Suppose we start with a hot gas and cool the system. Initially, the atoms will interact weakly. They will form structures that are inaccessible at low temperatures and sample a number of possible configurations. As we lower the temperature, only low-energy structures will be sampled. Eventually, if we cool slowly enough, we should quench out the lowest-energy structure or something close to this structure. We can simulate this process by performing a molecular-dynamics simulation using quantum forces, [29]. This is illustrated in Fig. 1.15 for a seven-atom cluster, [29]. We place seven atoms in a box and control their temperature by means of a hypothetical heat bath. We start off at a hot temperature and then cool the atoms slowly. (I should be clear here: “slowly” means that we consider a time frame of a picosecond or so. Of course, that is a short time experimentally, but it is a long time theoretically.)

As the seven silicon atoms interact, the initial bonds formed are weak, given the high initial temperature. Many structures are formed as the system is quenched. Eventually, we find a tetramer and a trimer, which form independently. In this simulation run, they merged to form a bicapped pentagonal cluster, which when formed resulted in a dramatic drop in the total energy of the system. How do we know if this structure is correct? Again, we can go to our silicon database. We can compare the theoretical calculations

to a number of experiments. For example, we can look at the vibrational modes of the cluster, optical absorption spectra, and photoemission spectra, [19, 29, 30]. Let me illustrate this for one example: a six-atom cluster of silicon, negatively charged (Si_6^-). We place this six-atom cluster in a heat bath and average the eigenvalue spectrum for this cluster as a function of time. This procedure yields a spectrum that can be compared to experiment, as indicated in Fig. 1.16.

The comparison is quite good, perhaps even better than we have a right to expect, given the simplicity of the calculation. The only “parameter” entering this calculation is temperature. We do not know what the temperature is experimentally, but we can make a reasonable guess, and the results are not overly sensitive to our guess. Here we have yet another example where we can test our theoretical tools on a new system (clusters), and we test our calculations against silicon.

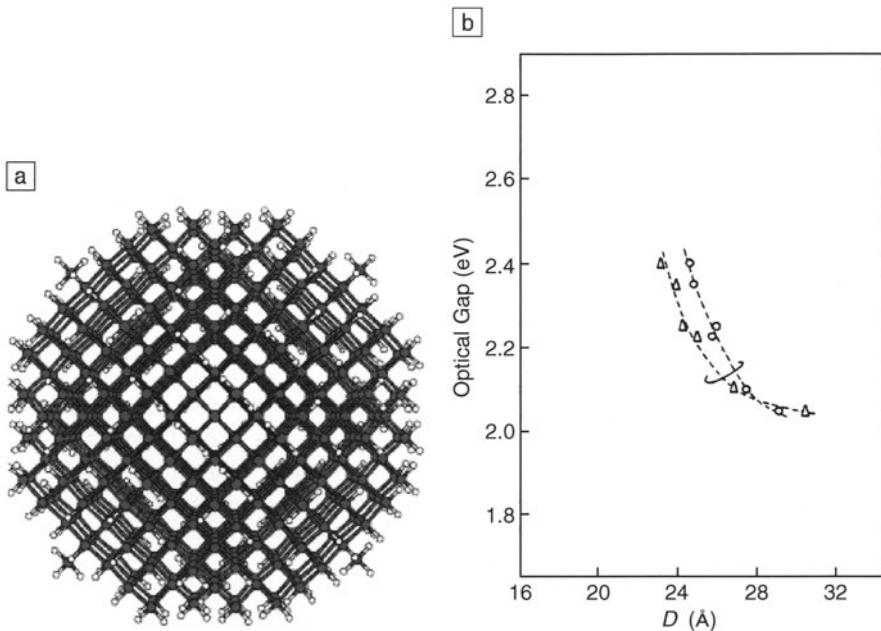


Fig. 1.17. **a** A hydrogenated silicon quantum dot. The solid gray circles are silicon atoms, and the open circles are hydrogen atoms. This quantum dot contains approximately 800 atoms. The silicon atoms are in a bulk-like environment, that is, each atom is fourfold-coordinated. The surface of this dot is passivated with hydrogen atoms. **b** The measured optical gap for such dots is illustrated as a function of the dot diameter D . As the dot shrinks in size, the optical gap increases. This change in gap size occurs because of quantum confinement. The measurements are from Furukaea and Miyasato, [32]

One more application: optical excitations in finite systems. These systems are of great interest because the size of the cluster can be used to tune the optical properties of the system.

I mentioned earlier in this presentation that the density functional theory was good for ground-state properties. However, some aspects of silicon do not lend themselves to ground-state properties. For example, optical properties are straightforward to measure but difficult to describe theoretically. In Fig. 1.17, I illustrate such an experiment, [32] for a hydrogenated silicon quantum dot – a ball of silicon 2–3 nm in diameter capped off with hydrogen atoms. The size of the observed gap in such quantum dots can vary from ~ 2 eV to ~ 2.5 eV; it is roughly twice the optical gap. It is widely accepted that confinement of the optical excitation within the dot increases the optical gap.

We can calculate the gap by solving for the Schrödinger equation as a function of time. We use the so-called time-dependent local density approximation (TDLDA), [33, 34]. We propagate wave functions in the presence of an electric field and find normal modes of excitation. Within this approach, we make one important approximation. We assume that the electronic coordinates respond immediately to the applied field (i.e., the adiabatic approximation). Does this approximation work? Is it reasonable? We look at silicon again.

Figure 1.18 shows a theoretical spectrum compared with experiment. In addition to the large hydrogenated silicon clusters (and quantum dots), we have included some small molecules such as silane and disilane. Experimental data are available for both of these molecules and the larger quantum dots. The agreement is rather striking, suggesting that at least in some favorable cases, TDLDA yields reasonable optical properties. This work remains in a formative stage, but shows much promise for examining the optical spectra of quantum dots and other nanostructures.

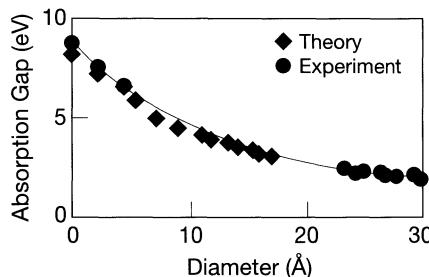


Fig. 1.18. Absorption (or optical) gap for hydrogenated silicon clusters and quantum dots. The theoretical data are from time-dependent local density functional theory, [33, 34]. The experimental data are from Furukawa and Miyasato for the dots, [32]. The cluster work is by Itoh et al., [35].

1.6 The Future

Albert Einstein said, “I never think of the future, it comes soon enough.”, [36] I think I will follow his lead. Predicting the future is not easy. However, I have two comments concerning where theory is going.

I said very little about advances in our computational tools, but not because they are unimportant. The advances in computational platforms over the past few decades have been breathtaking. For example, I have a PC on my desk in Minnesota comparable to the mainframe that I used for my thesis work in 1975. But if you posed the question to me as follows: would I rather have the ideas of the year 2000 and the computers of 1980, or the computers of 2000 and the ideas of 1980? The answer to this question is easy. Roughly 80–90% of my current work I could have done on those computers of the 1980s because of new ideas and concepts such as *ab initio* pseudopotentials, density functional theory, and new algorithms to solve the diagonalization problem, among others. In contrast, if you give me the computers of today with the ideas of 20 years ago, I am in poor shape. I could not do most of my current work.

In short, *ideas, not hardware*, have been the predominant driving force in the progress we have made. I have every confidence that ideas will continue to be more important, or at least as important, as hardware advances.

I will say one more thing about the future. I think *silicon* is and certainly will be the material of choice for theorists to test ideas with for many decades to come. On this prediction, I feel secure.

Acknowledgments

There are many people whom I really need to thank – indeed, more than space will allow. I have had not one, but three superb mentors: Marvin L. Cohen, James C. Phillips, and Morrel H. Cohen. I grew up with some great “intellectual brothers” such as Steve Louie, Jim Chadi, and John Joannopoulos. They remain good friends and confidants.

Also, I have had the pleasure of working with a number of excellent post-docs and students, many of whom can be found in the references included here. Of course, much of their work is reflected in this presentation.

Monetary support from the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Petroleum Research Foundation, and Exxon Research and Engineering Company is gratefully acknowledged, as is computational support from the Minnesota Supercomputing Institute.

References

1. See the following Web sites: PBS, “The First Silicon Transistor.”
<http://www.pbs.org/transistor/science/events/silicont1.html> (accessed September 2002); and Texas Instruments, “History of Innovation.”
<http://www.ti.com/corp/docs/company/history/sitrans.shtml> (accessed September 2002)
2. See the Web site at Intel on “Moore’s Law.”
<http://www.intel.com/research/silicon/mooreslaw.htm> (accessed September 2002), and on “Silicon Showcase: Breaking Barriers to Moore’s Law.”
<http://www.intel.com/research/silicon/> (accessed September 2002)
3. IEE/INSPEC science and engineering database home page.
www.iee.org/publish/inspec/ (accessed September 2002)
4. P.A.M. Dirac: Proc. R. Soc. London **123**, 714 (1929)
5. M. Born, J.R. Oppenheimer: Ann. Phys. **84**, 457 (1927)
6. D.R. Hartree: *The Calculation of Atomic Structures* (Wiley, New York 1957)
7. One of the first applications of the pseudopotential concept comes from E. Fermi: Nuovo Cimento **11**, 157 (1934)
8. J.C. Phillips, L. Kleinman: Phys. Rev. **116**, 287 (1959)
9. M.L. Cohen, J.R. Chelikowsky. In: *Handbook on Semiconductors*, vol. 1, ed. by W. Paul (North-Holland, New York 1982) p. 219
10. J.R. Chelikowsky, M.L. Cohen. In: *Handbook on Semiconductors*, vol. 1, ed. by P. Landsberg (Elsevier, Amsterdam 1992) p. 59
11. M.L. Cohen, J.R. Chelikowsky: *Electronic Structure and Optical Properties of Semiconductors*, 2nd edn (Springer, Berlin 1989)
12. M.L. Cohen, T.K. Bergstresser: Phys. Rev. **141**, 789 (1966)
13. J.R. Chelikowsky, M.L. Cohen: Phys. Rev. B **10**, 5095 (1974); Phys. Rev. B **14**, 556 (1976)
14. R.R.L. Zucca, Y.R. Shen: Phys. Rev. B **1**, 2668 (1970)
15. L. Ley, S. Kowalczyk, R.A. Pollak, D. Shirley: Phys. Rev. Lett. **29**, 1088 (1972)
16. J.R. Chelikowsky, N. Binggeli: Comput. Mater. Sci. **2**, 111 (1994)
17. L.W. Yang, P. Coppens: Solid State Commun. **15**, 1555 (1974)
18. W.E. Pickett: Comput. Phys. Rep. **9**, 115 (1989)
19. J.R. Chelikowsky: J. Phys. D **33**, R33 (2000)
20. P. Hohenberg, W. Kohn: Phys. Rev. **136**, B864 (1964)
21. W. Kohn, L. Sham: Phys. Rev. **140**, A1133 (1965)
22. M. Schlüter, J.R. Chelikowsky, S.G. Louie, M.L. Cohen: Phys. Rev. B **12**, 4200 (1975)
23. J.R. Chelikowsky, S.G. Louie (eds.): *Quantum Theory of Real Materials* (Kluwer Academic, Boston 1996)
24. M.T. Yin, M.L. Cohen: Phys. Rev. Lett. **45**, 1004 (1980)
25. K.J. Chang, M.M. Dacorogna, M.L. Cohen, J.M. Mignot, G. Chouteau, G. Martinez: Phys. Rev. Lett. **54**, 2375 (1985)
26. J.P. Perdew, K. Burke, Y. Wang: Phys. Rev. B **54**, 16533 (1996)
27. I. Stich, R. Car, M. Parrinello: Phys. Rev. B **44**, 4262 (1991)
28. J.R. Chelikowsky, N. Binggeli: Solid State Commun. **88**, 381 (1993)
29. N. Binggeli, J.R. Chelikowsky: Phys. Rev. B **50**, 11764 (1994)
30. N. Binggeli, J.R. Chelikowsky: Phys. Rev. Lett. **75**, 493 (1995)

31. L. Kronik, R. Fromherz, E. Ko, G. Ganteför, J.R. Chelikowsky: *Nature: Materials* **1**, 49 (2002)
32. S. Furukawa, T. Miyasato: *Phys. Rev. B* **38**, 5726 (1988)
33. M.E. Casida. In: *Recent Advances in Density-Functional Methods*, part I, ed. by D.P. Chong (World Scientific, Singapore 1995) p. 155
34. I. Vasiliev, S. Ogut, J.R. Chelikowsky: *Phys. Rev. Lett.* **86**, 1813 (2001)
35. U. Itoh, Y. Toyoshima, H. Onuki, N. Washida, T. Ibuki: *J. Chem. Phys.* **85**, 4867 (1986)
36. See “Einstein Links” on the PBS Web site.
<http://www.pbs.org/wgbh/nova/einstein/links.html>
(accessed September 2002).

Part I

“The” Crystalline Bulk Semiconductor Silicon

2 Silicon: the Semiconductor Material

W. Heywang, K.H. Zaininger

2.1 Introduction

Fifty years of silicon for semiconductor device applications is the milestone at which this series of articles has been written, which will present the many-faceted development of all the technologies that are connected with it, their present status, and recognizable future trends. The individual articles will cover topics such as:

- growth of single crystals and its reproducibility in industrial applications
- polycrystalline and amorphous silicon
- epitaxial technologies and thin films
- crystal defects, impurities, and doping
- various processes for micro- and nano-structuring
- materials requirements from the vantage point of the users in the fields of microelectronics, power electronics, optoelectronics, and micromechanics
- interfaces to other materials such as III-V compounds, as well as the whole area of bioelectronics.

Since this series is about silicon it is taken as self-evident that all these contributions will emphasize the material aspects.

This undertaking is justified and useful because silicon has, like no other material, dramatically changed our world. Especially, the whole of information and communication technology would have developed completely differently without the availability of silicon. Thus, we can with all justification talk of a silicon era, just like one talks of a stone, copper, bronze, or iron age, where a specific material that predominantly characterized the advancements made during that time was chosen for the name of that era. Even more than with iron and steel, we have to deal here with a multitude of individual technological advances which ultimately made the material as we know it today. The variety of the topics mentioned above attests to this. Last but not least, it has to be mentioned that silicon, because of the intense technological interest in it, belongs to the small group of solids that have been exhaustively studied and where, during its study, significant contributions to materials science have been made.

We were invited to write an introductory paper to this series of articles because we belong to those few people that have lived through this development

from its very first beginnings – originally as technical-scientific researchers and later in responsible managerial positions. One of us (WH) was at Siemens, initially involved in the materials research and subsequently in device application. The other (KZ) was at RCA Laboratories in Princeton, NJ, USA, where he devoted his research efforts solely to silicon device technology. We have the indisputable advantage of a personal understanding of the various problems and an involvement in the sometimes dramatic happenings. On the other hand, there exists also the danger of a one-sided opinion. We believe that the advantage outweighs the disadvantage and that we can be rather objective; and so we have accepted this task.

In our discussion we want to show that silicon – during all of its developmental stages – in no way ever looked clearly to be the obvious solution to the various problems as we so easily see it in retrospect today. On the contrary, silicon's march forward was often an adventurous path that had many individual successes, but also erroneous paths and fallacious assessments, as is common in all research- and development activities.

The silicon semiconductor technology formed the basis for the development of the information society, a society which is characterized by the mental achievements of humankind. This partially virtual world – broken down into bits or built up from them – seems sometimes only a product of the human brain, a world made by humans. But still, where would this world be without silicon?

Silicon does not appear as a free element in nature, because of its high chemical affinity, especially for oxygen. It can only be produced by chemical reduction. Is it then also a product of man's creativity, custom tailored for his purpose? Or is it – with its special properties – still nothing else but a wonderful present of nature?

2.2 Early History

When, in this contribution to *Silicon*, we talk about making silicon available for semiconductor applications, it is not just for history's sake. No, we also want the readers to understand how this development was possible, what parts were targeted goals, and what was simply serendipitous discovery? Here we immediately come to the age-old question of humanity: Can man create his own world – as often described or shown in some science fiction of our time – or is he still just embedded in this nature and creation that was given to us as a present?

Despite some fundamental work carried out in the 1920s and 1930s on the band structure, and theoretical attempts concerning rectifier effects and several patents concerning the unipolar transistor in that period, the real semiconductor era began only with the proof of the bipolar transistor effect in germanium by Bardeen and Brattain at the end of 1947 [1]. On the basis of the rapidly increasing industrial interest, the procedures for crystal growth,

purification, and doping – still important even today – were then developed in the 1940s and 1950s, but initially just for germanium.

At the same time, an increased interest in other semiconductor materials appeared. Were there perhaps materials other than germanium that would be even more suitable for specific applications? This question could only be answered by first understanding the transistor effect. The extremely high carrier mobilities in germanium, about 10^3 higher than those of oxide semiconductors, were fascinating here. Achieving even higher mobilities was then one of the goals of the pioneering work in the field of III–V compounds by Heinrich Welker and his research team at the beginning of the 1950s [2].

Research on silicon began at about the same time, but rather in the background. Sensational discoveries and/or advances were hardly expected unless one considered the proof of the semiconducting character of silicon – still in doubt in the 1940s – as such. The metallic shine and relatively high conductivity of the (highly impure) samples, similar to that in the so-called hard metals, led to this erroneous conclusion. For example, in the 1953 edition of Linus Pauling's book *General Chemistry* [3], silicon is still called a semimetal. On the other hand, Pearson and Bardeen [4] had discovered in 1949 the high-temperature transition to intrinsic conduction and, thus, presented the proof of the semiconducting character of silicon. The samples used were, however, not single-crystal, so that the conductivity below this transition did not show the typical temperature dependence of extrinsic semiconductor conductivity. This caused some irritation amongst the various researchers [5, 6].

The experiments of Pearson and Bardeen showed that the bandgap was 1.12 eV. The carrier mobility could also be determined at the beginning of the 1950s by use of Hall and drift measurements. With a value of $1200 \text{ cm}^2/\text{Vs}$ for electrons and $300 \text{ cm}^2/\text{Vs}$ for holes, it was about a factor of three lower than that of germanium. All this did not seem to be very exciting, and thus, there were only a few researchers who devoted themselves to silicon and its crystal growth.

Because of the increasing importance of microwaves in the mid-1940s various semiconductor detectors, based partially polycrystalline silicon, were used in microwave applications to replace microwave tubes because of their small dimensions and high cut-off frequencies. So even at that time, silicon was making inroads into communication technology. This was also where the newly discovered transistor was of special interest, because a higher carrier mobility is a definite advantage for achieving higher cut-off frequencies. So, a theoretical comparison – since the transistor effect in silicon was not yet experimentally proven at that time – clearly indicated the advantage of germanium for use in communication technology, and the limit on the thermal stress of 80°C , caused by the bandgap of only 0.7 eV, was considered acceptable. Only in cases of elevated heat production were these limitations serious, but it was hoped that they could be overcome by use of other semiconductors such as the III–V compounds already mentioned or by silicon. The high

degree of adaptability, the high carrier mobility (in GaAs, for example), and a lower melting point favorable for processing, were all strong arguments for the III-V compounds. In silicon, on the other hand, one had to deal only with one kind of atom, which eliminates various kinds of lattice defects and guarantees a good lattice thermal conductivity.

So much for the theoretical considerations. In the reality of the technological world of the early 1950s one was far away from a general use of silicon as a semiconductor material. As already mentioned, because of silicon's high chemical aggressiveness at the elevated temperatures required for its preparation, all silicon samples were highly impure. Their level of purity was, as Pearson and Bardeen's samples showed, in the range of several ppm. Despite this, research groups, especially that of G.K. Teal at Bell Laboratories, were not discouraged. Teal devoted himself, in addition to his main task of germanium, to silicon crystal growth until he went to Texas Instruments in 1952. But even there he built up a silicon research team which had the goal of developing power transistors. The move to TI required time, and that significantly reduced the original advantage he had over possible competitors.

2.3 Competition and Cooperation in the Silicon Race

At the beginning of the 1950s, but at first entirely in a clandestine mode, another competitor came onto the scene: the Siemens Group, of which WH was a member. Despite entering late, Siemens definitely wanted to compete in the important, newly blossoming area of semiconductor physics and technology, having already its own manufacturing facility for selenium rectifiers, located in its power group; there was also a production facility for transistors, using its own experience base in high-frequency rectifiers and following Bell Laboratories as far as germanium was concerned. With regard to additional research, it had one of the best-known semiconductor pioneers, Walter Schottky, within its own ranks. For new semiconductors, Heinrich Welker, with his pioneering ideas and experience in III-V semiconductors, was appointed as department head for the new research laboratories of Siemens-Schuckert.

In this technological environment, the materials research laboratory of Siemens and Halske also entered the semiconductor area, and started in 1951 to work mainly on trial-and-error experiments on silicides such as Mg_2Si . The purification of silicon was included as a prerequisite in these fundamental studies. To achieve this, two approaches were pursued: on the one hand the classical metallurgical preparation of pure silicon powder via magnesium or zinc reduction of pure quartz (the so-called B-process), and on the other hand the reduction of $SiHCl_3$ by hydrogen in an electrical discharge according to the method of van Arkel. This second approach was carried out in collaboration with Wiberg and Amberger of the University of Munich and was called the A-process.

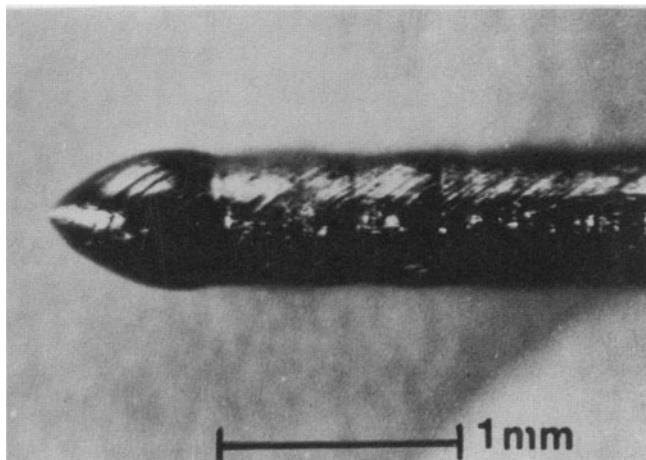


Fig. 2.1. First “high-purity” silicon rod from Siemens grown by the so-called A-process in 1953

While experiments were being carried out with this latter procedure a great surprise happened, namely, on one of the electrodes a thin silicon rod had grown (Fig. 2.1) which, in its single-crystal part, showed a purity that was orders of magnitude higher than that of all samples obtained by the B-process [7].

The specific resistance was 20 ohm cm , whereas the others had less than 0.1 ohm cm . The breakdown voltages of needle electrodes were about 100 V and exhibited the polarity of an n-type semiconductor. So, with one step, the purity was increased by at least three orders of magnitude and was in the sub-ppm region. These first results could be reproduced, and subsequently the purity, the diameter of the needles, and the size of single-crystal areas could be improved even further. Our group could compare itself now with its international colleagues, and concentrated its efforts on the crystal growing and purification of silicon *as its only goal*.

Fortunately, the aforementioned laboratory for power rectifiers, under the leadership of E. Spenke in Pretzfeld, still working on selenium, also joined the silicon work. Spenke had, at that time, already recognized the importance of silicon for power applications and stated, “Silicon is, just like germanium, an elemental semiconductor and therefore does not have many of the defects that are possible with compound semiconductors because of imperfect stoichiometry and elemental dislocations. It has a sufficiently large band gap and a carrier mobility that is definitely acceptable for power applications. Thus, we will bet on it.” It is obvious that, with this decision, the internal competition with Welker’s research group working on III–V compounds for the same applications was preprogrammed.

In Spenke's remarkable prognosis, made in 1953, one important characteristic of silicon is understandably not mentioned, namely the long carrier lifetime. This is a result of the specific band structure of silicon, which it shares with germanium, but was not yet known at that time, namely that the minimum of the conduction band – contrary to the original assumption – does not occur at the wave vector $k = 0$ but, rather, near the edge of the Brillouin zone. This prevents a direct optical recombination of electrons and holes, which would occur in a time on the order of magnitude of 1 microsecond (as, for example, in GaAs). This, of course, prevents, on the one hand, the technological use of silicon in the area of active light generation but allows, on the other hand, the achievement of long carrier lifetimes (just as in germanium) for high purity and perfect crystallinity. This gift of nature was a crucial (at that time of course unknown) precondition for the discovery and development of the bipolar germanium transistor, because it is the means for achieving the required diffusion lengths of minority carriers. High diffusion lengths are important factors for the I–V characteristics not only of bipolar transistors but also of all power devices such as rectifiers and thyristors.

With these comments, we have got a little bit ahead of ourselves, because these fundamental characteristics of the band structure were only discovered step by step towards the end of the 1960s beginning with the work of Herman et al. [8]. So, let us return again to the year 1953 and to the then newly discovered levels of purity in the silicon samples produced by the von Arkel method. These results were quite encouraging but the limitation of sample diameters to a few mm made the samples unsuitable for practical application, especially in the power area. The question now arose of whether or not the electrical discharge, which was concentrated on small electrode areas in the van Arkel method, really was the key to the high-purity effect, or could the thermal reduction of the SiHCl_3 /hydrogen mixture on a hot, glowing silicon surface (“CVD, chemical vapor deposition” as it is called today) be sufficient? In order to answer this question, thin needles were heated in a specially developed reactor by sending current through them and could thereby be covered with a thick layer, which was, however, polycrystalline. To be able to assess the purity of these polycrystalline samples by resistance measurements, they had to be transformed into crystalline material, preferably into a single crystal. This had to be achieved without addition of more impurities, as would be caused for example by touching of the wall of the crucible. To achieve this, the vertical zone melting method was invented in 1952 almost concurrently by K.H. Theuerer at Bell Laboratories and K. Siebertz and H. Henker at Siemens. In this procedure a molten zone, produced by high-frequency (HF) heating, is carried by the original ingot or, as in the case of Fig. 2.2, by the already solidified new monocrystal below the glowing molten zone.

The molten zone is held stable by surface tension, as well as by the electromagnetic forces of the HF heating. The new crystal then grows out of the molten material. This zone melting method was further improved in Spenke's

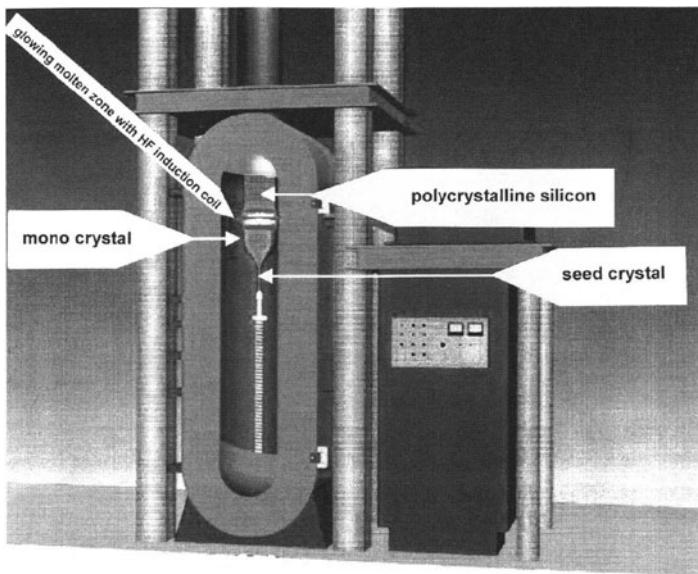


Fig. 2.2. Vertical zone melting (courtesy Wacker Chemtronics)

laboratories so that the newly forming crystal rod was not rigidly connected to the original rod that was to be melted. Using different pulling speeds for the upper and lower rods, one could arbitrarily achieve different diameters for the newly formed part of the crystal. This was used in the example of Fig. 2.2 for the transition from the small diameter of the seed monocrystal to the desired final diameter of the growing rod. In the early stages Spenke's group used this procedure for the growing of thin rods, which could then be thickened by the CVD method. A cross section of such a thickened sample can be seen in Fig. 2.3.

By this means, a closed procedure – independent of the A-process – had been achieved for wall-free production of high-purity single-crystal silicon from the gaseous phase, which, after a number of additional technological improvements, was also suitable for mass production. With this complete method (now simply called the C-process), it was possible to produce silicon rods with a diameter of a few cm and a length of more than 1 m. The maximum diameter of these rods was, however, restricted owing to the physical limitations set by the vertical zone melting itself. Thus, because of the ever-increasing demand for larger-diameter silicon wafers, and after some years of further technological development concerning oxygen and other contaminants as well as crystal quality, the vertical zone melting procedure was pushed aside again by the classical pulling from a crucible according to the method of Czochralski. A picture of such a crucible-grown rod when it just has been drawn out of the crucible is shown in Fig. 2.4.

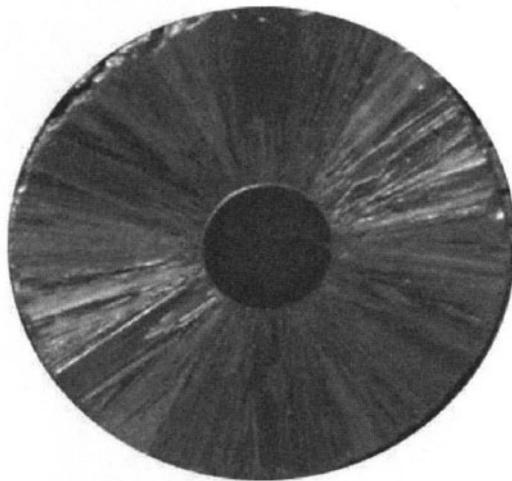


Fig. 2.3. Cross section of a thickened silicon rod. The monocrystalline center is surrounded by a polycrystalline CVD layer



Fig. 2.4. Crucible-grown modern dislocation-free silicon monocrystal (courtesy Wacker Chemitronics)

Nevertheless, the development of the vertical zone melting process was, at that time, an absolutely necessary step for the industrial breakthrough of silicon as a semiconductor material. The complete C-process had opened the way to industrial mass production.

During the time of its development, vertical zone melting was the key for thoroughly investigating the purity, properties, and crystal quality of all our silicon samples. This may be illustrated by the following story. When increasing the diameter of thin crystals by CVD layers we found, after deposition and transformation into single crystals, an unexpectedly high conductivity of the final single-crystal rod. This was found to be due to a donor concentration that was close to 1 ppm. However, tests in an A-reactor using the same SiHCl_3 samples had shown a prevalence of acceptors with a concentration of about 0.01 ppm. Where did the high donor concentration come from? Were the donors impurities from the reactor? Were the samples contaminated by inadequate handling? Was this a problem of the deposition process itself? A lot of questions had to be answered simultaneously in the sub-ppm regime. A first positive result was that the impurities could be removed by multiple zone-pulling. But that was obviously cumbersome.

What kind of impurities were these anyway? This question could not be answered easily, because of their small concentration. So, normal chemical-analysis methods were way too insensitive. A suspicion concerning phosphorus could not be confirmed by a radiotracer method with neutron activation. The results were confusing. Furthermore, we measured the segregation coefficient for the desegregation during the zone melting. It did not fit anything. Thus, one talked about the donor X. Only the joint efforts of Honrath and Ziegler [9] produced clarity: during the radiotracer analysis, the neutron radiation not only activated the phosphorus but also created Si-31 from the Si-30 isotope which is contained in natural silicon. This transforms, through a gamma-process, into phosphorus. This additional phosphorus was then activated in a second step and consequently, depending on radiation dose and exposure time, resulted in erroneous results. In the zone-pulling procedure, on the other hand, one had to take into account not only the segregation coefficient of phosphorus between the molten material and the silicon crystal but also evaporation, which depended on the diameter of the molten zone.

With this knowledge – simple only *a posteriori* – it was possible to identify phosphorus as the critical n-impurity. And then it became possible to reduce its concentration and thereby solve the problem.

The story of the donor X may be considered representative of the problems we were confronted with in this early phase. Penetrating this region of ultra-high purity required thinking in new dimensions. The problems could only be solved by interdisciplinary cooperation: novel chemical means and equipment for the purification of the starting chemicals; measuring by overcoming oxide barriers and blocking layers; clarification of donor-acceptor compensation; radiotracer analysis; thermodynamics of crystal growth, including phase

diagrams; interactions between the various contaminants and with lattice defects; etc. Thus, it is obvious that a close cooperation had to be developed between the different Siemens groups on the one hand and the suppliers of the chemical substances on the other hand. This mainly concerned the metal-organic compounds of the Si–H–Cl system. There was a continuous exchange of samples and data between our different groups on the one hand and Wacker (later on Wacker-Chemtronics) on the other hand. This cooperation had, at first, developed open-mindedly but then it was formalized and, finally, it led to a licensing agreement for the entire production and test process.

Two special results of this early company cooperation should be mentioned:

- The boron problem: boron is an acceptor dopant still in use for device production. In contrast to phosphorus, it cannot be removed by zone refining or evaporation, because of its chemical similarity to silicon. Thus we had to rely entirely on Wacker's purification process and reliability.
- The carbon problem: chemically, carbon is homologous to silicon and does not act as a donor or acceptor; it cannot be detected by resistivity measurement. But, through the formation of SiC precipitates, it prevents monocrystalline crystal growth. These SiC precipitates were the result of $\text{SiHCl}_2\text{CH}_3$ contamination of SiHCl_3 . Again, we depended here entirely on the capability and quality control of Wacker.

There was also some work on silicon at IBM in Schwuttke's group. However, it did not result in many pivotal advances in the field of very pure silicon. The contributions were more along the lines of crystal defect analysis by x-ray topography.

These few examples from the early cooperative efforts may underline again that the development process described above was an expedition into the unknown that could have easily gone astray. But high-purity, perfect silicon crystals are not a purpose in themselves; they represent a crucial precondition for the proper functioning of semiconductor devices, which, in turn, are a result of proper doping, structuring, contacting, etc. All of this forms a separate set of problems, which had to be gotten under control through relevant research and development. These topics will be treated in separate contributions within the framework of this series, so that we do not have to go into details in this introduction.

2.4 Initial Device Applications

It seems significant to us that the Texas Instruments and Siemens groups, which worked independently on silicon, had also set themselves different goals. The main goal of Teal was the development of bipolar silicon transistors. He showed initial results at the 1954 IRE Conference in Dayton during his presentation "Some recent developments in silicon and germanium materials and

devices.” But, as he stated himself in 1976 in his survey article [10], he could hardly overcome the existing reservations concerning the reproducibility and stability of silicon transistors. In addition, there was a lower cut-off frequency due to the lower carrier mobility for comparable geometries. This did not, however, deter him from continuing to pursue this approach, especially for military applications that had higher power requirements.

Concerning the application of silicon in communication and information technology Siemens (at that time Siemens & Halske) remained sceptical despite the advances in the development of the highest-purity silicon – following the philosophy of Bell Laboratories. Only at Siemens-Schuckert (responsible for power technology) did Spenke and his group devote themselves consistently to the development and technical improvement of silicon for commercial application in high-power rectifiers in order to replace the old selenium technology. This new approach, including device technology, was ready in 1956 for application and was used at first for power rectifiers in Siemens locomotives [11]. Obviously, the total internal Siemens power-electric market was immediately available for the application of these rectifiers – an advantage not to be underestimated. Contrary to other companies, Siemens did not pursue a policy of a monopoly in the silicon market, but rather gave licenses for the total high-purity silicon manufacturing process not only to Wacker in Germany but also to other chemical companies, mainly in the USA and Japan. These licensing agreements also included cooperation, so that further improvements could be exchanged, and some kind of worldwide cooperation developed. This had the consequence that one can say today that the total world production of silicon semiconductor material was initiated by the so-called Siemens Process. An example of a modern silicon rod can be seen in Fig. 2.4 and should be compared with the mini-rod of half a century ago (Fig. 2.1).

As far as commercial power applications were concerned, Siemens obviously did not stop with the development of silicon rectifiers; transistors and thyristors were also developed. Here, they encountered the same problem of instability in the I–V characteristics that Texas Instruments had already seen. One of the important concerns was related to the trapping problem, i.e. the time-limited presence of charge carriers at defect sites. Bell Laboratories had already encountered this problem in 1953 during the examination of their silicon [12]. Was this trapping effect possibly due to “natural” defects that were intrinsic to silicon and therefore unavoidable, for example, lattice defects? We also attacked this problem and thoroughly examined various trapping effects in the silicon that was produced in Spenke’s laboratories and also elsewhere [13, 14]. The most important result of those investigations was that the trap density in the silicon produced by the Siemens process was below $10^{11}/\text{cm}^3$ and, therefore, caused no problem at room temperature. The reason for that was the absolutely wall-free preparation of this silicon – an advantage not planned but nevertheless crucial.

2.5 MOS Technology and Integration

Despite all these above-mentioned encouraging results at Texas Instruments and Siemens, the breakthrough for the use of silicon instead of germanium in the communication and information area was not at all guaranteed yet. Even at the end of the 1950s, higher cut-off frequencies seemed to be the determining factor. The higher carrier mobilities found in Ge and GaAs were far more attractive. The breakthrough of silicon came only with two innovations which were unexpected to most semiconductor specialists:

1. The discovery that thermally grown SiO_2 films passivate (stabilize) the surface of the silicon substrate was the turning point in semiconductor technology resulting in the change from germanium to silicon, since germanium surfaces could not be stabilized in a similar manner. This passivation was due to a significant reduction of the surface states (dangling Si surface bonds) by forming Si-surface–O bonds and made possible the control of charge carriers via the oxide layer. The low density of surface states in Si–SiO_2 structures opened up the way for developing the metal–oxide (SiO_2)–semiconductor (Si) field-effect transistor (MOSFET), which, in its basic concept, utilizes ideas that were already known in the 1920s (e.g. [15]), but which were at that time miles away from being able to be realized. In the development of this technology, which is today the basis for microchips, especially those used in computers, the properties of the SiO_2 film as a dielectric and mechanical layer also played an important role.

The first realization of an MOS transistor was achieved in 1960 by Kahng and Atalla at Bell Laboratories [16] and was followed in 1961 by Zaininger at RCA Laboratories [17]. But it still required a large amount of additional developmental work until the MOS transistor technology was under control and reproducible. This will be treated in detail in one of the following articles. Therefore we shall limit ourselves to simply pointing out that, before the development of the MOS technology, it was a downright statement of faith that stable semiconductor devices had to avoid or at least to fight the detrimental influence of the surface.

2. The second innovation is also related to the surface. It is the concept of device integration within a single silicon chip, which, at its beginning, seemed to many rather limited in its possibilities. One has to remember that the yield for single transistors at that time was about 30%. Thus, by simple deduction one would expect the yield for two hard-wired transistors to be 0.3×0.3 , i.e. about 10%, for three transistors about 3%, and so on. It was Kilby's brilliant mind [18] which realized that it is really the metal wire connection that is the main source of failures and that avoiding it by "integrating" could lead to higher yield and reliability. Indeed, he built the first functioning integrated circuits, initially a phase shift oscillator and then a few flip-flops, in the summer of 1957 by connecting

about 25 transistors that were on a 5 inch germanium wafer. But who at that time would have seriously thought of the high levels of integration that are characteristic of today's microelectronics? These developments would not have been imaginable without the specific masking properties of the silicon dioxide layer. Since this question, together with the many required research topics and technological developments, will be discussed in detail in the following articles, we shall limit ourselves to just this one question. Which properties of silicon are crucial for the role of the silicon oxide layer?

One property is pretty obvious. It is the high purity. Only on a pure material is it possible – provided that additional subsequent procedures are also clean – to grow a defect-free oxide in which, for example, no ions are contained that could move when an electric field is applied to the oxide and thereby change the I–V characteristics. Another property, at least as important as the first, is the chemical stability of the initial oxide layer. A clean, undisturbed silicon surface is unstable, as was shown, for instance, through investigations with LEED (low energy electron diffraction) on free silicon surfaces in high vacuum [19]. For example, a normal (111) silicon surface clearly shows the sixfold symmetry of the uppermost layers, as expected from this crystal structure. However, this changes during annealing in ultra-high vacuum into a much more complicated symmetry in which the dangling bonds satisfy each other (Fig. 2.5).

But this “pure” state is only stable in high vacuum and eagerly attracts atoms of other elements. Important for our considerations here is the fact that these are preferably oxygen atoms, which then can form an initial oxide layer. This stable and atomically dense first layer is a crucial advantage of silicon and does not exist in other competing semiconductor materials. This is possible because of a fortunate situation, namely that the distance between the silicon atoms in the Si–O–Si bond coincides with the distance between two silicon atoms in the diamond structure of the basic lattice. Only this makes the growth of the dense and defect-free initial oxide layer explainable and the avoidance of undesirable surface and interface states possible. The atomically dense transition of the monocrystalline silicon substrate to the grown-on SiO_2 layer is shown in Fig. 2.6.

In spite of the atomic fit, the oxide layer is amorphous because of the differing crystal symmetries of silicon and silica. Provided that we are dealing with a perfect stabilization of the silicon surface by oxygen saturation of all dangling bonds, to suppress all undesirable interface states, the only requirements which have to be fulfilled by the follow-up silica layer are stability and atomic density. This is necessary so that no contaminating ions are able to penetrate or migrate, since even minor migrations would cause instabilities of the device characteristics.

Furthermore, we recognize that the amorphous silica layer on top of the stabilized interface can be replaced by other layers provided the stability

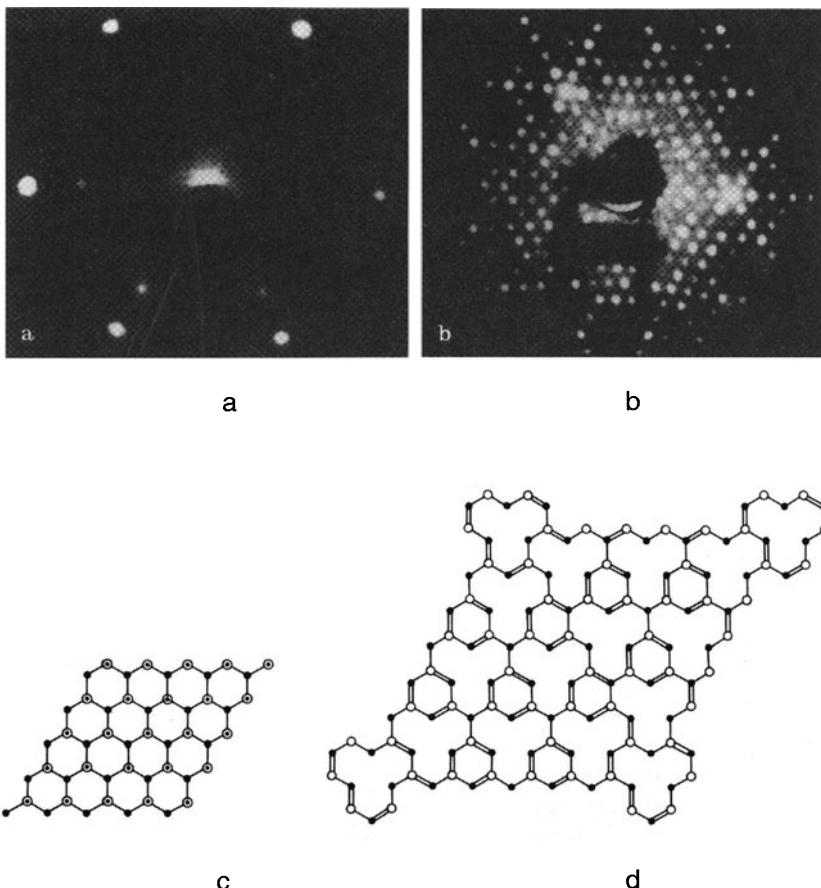


Fig. 2.5. Silicon surface, crystal structure. (a) LEED pattern of oxygen-stabilized (111) surface; (b) LEED pattern of oxygen-free (111) surface; (c) crystal structure of a (111) silicon surface with vertical dangling bonds, which have to be saturated by atoms of the next (not shown) layer; (d) self-saturation of dangling bonds, only stable in a vacuum that is better than 10^{-9} torr: \circ atoms of the top layer with dangling bonds, \circ atoms of the top layer without dangling bonds, \bullet atoms of the second layer from top

requirements mentioned above are fulfilled. This degree of freedom is used, for example, in the case of Si_3N_4 passivation and wherever else silica is to be replaced by low- k or high- k materials for improving speed or integration density in modern IC technology.

In summary, through this interface stabilization, nature has helped us again in the development of IC technology, including the modern varieties we are working on today. This is an effect which nobody thought of when the work on the use of silicon for semiconductor purposes began, especially at

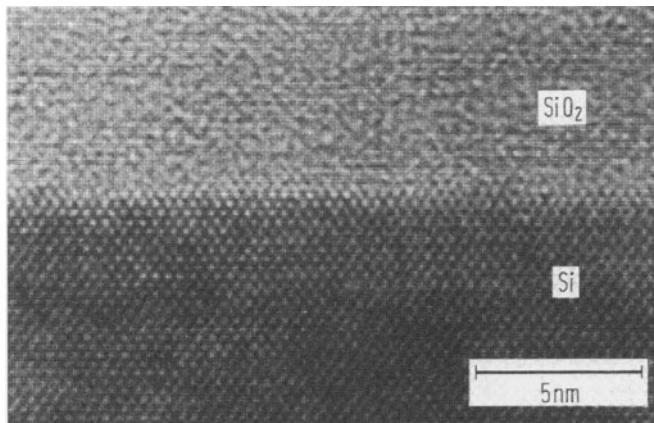


Fig. 2.6. Interface between silicon substrate and SiO_2 surface layer in atomic resolution

a time when there were so many other problems connected with the always present oxide layer during measurements and contact making.

Finally, the following fact, which was crucial for the victorious advance of silicon into information technology, has to be pointed out. Advanced MOS Technology and very large-scale integration (VLSI) – both based on the extraordinary properties of the $\text{Si}-\text{SiO}_2$ system – opened the way for a cost-effective digital technology, which then, in turn, opened the way for the necessary mass market in silicon integrated circuits for use in information technology.

2.6 Conclusion

With these remarks, which already reach into the heart of the silicon era, we want to conclude this review of the pioneering times of silicon, *the semiconductor material*. We hope that this description of the whole development of pure, single-crystal silicon makes three things clear:

1. Silicon, with all its positive basic attributes, had first to be made readily available for large-scale common use. This was an often difficult path into the unknown, where only the human pioneering spirit had a chance to overcome all the difficulties that were encountered. It certainly was not a straightforward development – as it might appear to someone in retrospect and as extrapolations such as Moore's Law might make one believe – but rather one that was characterized by many, often extremely complex individual developments. It was an outstanding achievement of the human spirit in research, as well as of interdisciplinary cooperation between material researchers, device developers, designers, and technologists.

2. Most importantly, it is obvious that a uniform, high-purity, perfect single-crystal piece of silicon material alone would be of extremely limited technological use. It is only when this material is suitably altered and structured through controlled, reproducible processes and then made into useful devices that it becomes valuable. All of these requirements could not have been achieved were it not for the fact that nature provided us with an extraordinary gift, through suitable physical properties and constants, etc. – an abundance of wonderful, often crucial properties and characteristics of silicon, silicon dioxide, and the Si–SiO₂ interface that, together, make modern integrated silicon technology possible at all. Let us just quickly enumerate the most important properties and characteristics:

2.6.1 Silicon

- Abundant: easy to obtain, low cost.
- Single crystal: with ever larger rod diameters (30 cm). Defects can be eliminated or selectively utilized for advantage.
- Not brittle: can easily be handled and is an excellent mechanical substrate for individual devices and integrated circuit chips.
- Adequate thermal conductivity to take away the electrically generated heat in chips.
- Can be microstructured by a combination of suitable optical and chemical methods (lithography), even breaking through the 0.1 micron barrier.
- Thin crystalline silicon films with different electrical properties can be grown onto silicon substrates via epitaxy.
- Thin crystalline silicon films with various electrical properties can be grown onto insulators (sapphire, etc.) to provide improved isolation and speed, and lower capacitance.
- Thin crystalline germanium films and, probably, novel films of III–V compounds containing quantum dots, offering different electrical and optical properties, can be grown onto silicon substrates via chemical vapor deposition or molecular-beam epitaxy.
- Buried thin films of SiO₂ can be created under the silicon surface by oxygen ion implantation and subsequent annealing (SIMOX structures).
- Has a very useful energy gap (1.12 eV).
- Conductivity can be tailored (n-type, p-type, value) by doping using diffusion and/or ion implantation.
- As an elemental semiconductor, it does not have the multitude of materials problems and chemical behavior that compound semiconductors have.
- Annealing works very well.
- Carrier mobility is good for both electrons and holes (important in CMOS circuits).

- Carrier lifetime for both electrons and holes is good because of special band structure properties and low density of traps (important for bi-polar devices).
- Not light-sensitive (stable operation of devices under various light conditions).

2.6.2 Silicon Dioxide [20]

- Can be thermally grown as a native oxide by a simple, inexpensive and reliable (oxidation) process.
- Can be deposited via chemical vapor deposition and other methods.
- Is stable up to very high temperatures (important for annealing).
- Films can be very thin (100 Å) (necessary for ultrasmall MOS devices).
- Acts as a chemical barrier during etching of selected silicon areas,
- Can act as a diffusion barrier for certain materials, especially most of the common dopants.
- Acts as a barrier during ion implantation.
- Is chemically stable but can be microstructured by a combination of suitable optical and chemical methods (lithography), even breaking through the 0.1 micron barrier.
- Metal patterns, deposited on it by various methods, adhere very well.
- Is mechanically strong and can act as a protective layer (physical and ionic protection).
- Can be polished to planarize the surface.
- Is transparent.
- High electrical breakdown strength.
- Useful dielectric constant (but here there is a need for new insulating materials to replace SiO_2 in certain areas: high- k materials for the gate and low- k materials for insulating the wiring).

2.6.3 Si– SiO_2 Interface [21]

- Has an extremely low density of interface states when properly prepared.
- Very stable.

Had this been different – either through a quirk of nature or a dispensation of providence – the development would have taken a different path, and the enormous advances we see today, especially in microelectronics and information technology, would hardly be imaginable.

3. As outlined in item 2 above, silicon is a unique gift of nature, so that we are justified in speaking of *the* semiconductor material. It cannot be supplanted in its importance to our technology and will continue to dominate the core of semiconductor electronics. Important complementary technologies can only be expected in areas where silicon encounters its

natural limits, for instance in optoelectronics, large display technologies, sensor technology, or modern bioelectronics. In these areas one also needs new materials and material combinations which might partially contain silicon.

However, significant new applications may still be discovered for this unique gift of nature, making it then one of the most significant if not the most significant material in the world today.

Acknowledgments

We are indebted to all long-time friends who have helped us and contributed to this review with their own experience. We have to thank Dipl.Phys. Remigius Pastusiak for digitizing and improving the contrast of old Siemens pictures.

References

1. J. Bardeen, W. Brattain: Phys. Rev. **75**, 1208 (1949)
2. O. Madelung: *Physics of III-V-Compounds* (Wiley, New York 1964)
3. L. Pauling: *General Chemistry*, 2nd edn (W.H. Freeman, San Francisco 1953)
4. G.L. Pearson, J. Bardeen: Phys. Rev. **75**, 865 (1949)
5. E.M. Conwell: Proc IRE **40**, 1327 (1952)
6. W. Heywang, M. Zerbst, F. Bischoff: Die Naturwissenschaften **42**, 301 (1954)
7. W. Heywang: Internal Siemens Report (7.8.1953)
8. F. Herman, R.L. Kortum, C.D. Kuglin, R.A. Short: New studies of the band structure of the diamond-type crystals. In: *Proc. 8th Int. Conf. Phys. Semicond.*, Kyoto (1966) pp. 7–14
9. G. Ziegler, M. Honrath: Internal Siemens Report (3.6.1958)
10. G.K. Teal, K. Storcks: IEEE Trans. Electron. Devices **23**, 621 (1976)
11. E. Spenke: Report at the Garmisch Conference, Oct. 1956. In: *Halbleiter und Phosphore*, ed. by M. Schön, H. Welker (Vieweg, Braunschweig 1958) pp. 630–640
12. J.R. Haynes, J. Hornbeck: Phys. Rev. **90**, 152 (1953)
13. W. Heywang, M. Zerbst: Z. Naturf. **14a**, 641 (1959)
14. M. Zerbst, W. Heywang: Z. Naturf. **14a**, 645 (1959)
15. J.E. Lilienfeld: US Patent No. 1745175 (1926)
16. D. Kahng: IEEE Trans. Electron. Devices **23**, 655 (1976)
17. K. Zaininger: RCA internal report (1961)
18. J.S. Kilby: Miniaturized Electronic Circuits. US Patent No. 3,138,743 (1959/1964) filed February 1959, issued 23 June 1964
19. J.J. Laner: Low energy electron diffraction and surface structural chemistry. In: H. Reiss: *Progress in Solid State Chemistry*, vol. 2 (Pergamon Press, Oxford 1965) pp. 26–90
20. A.G. Revesz, H.L. Hughes: The Structural Aspects of Noncrystalline SiO₂ Films on Silicon, A review, J. Non-Cryst. Solids **328** (1–3), 48–63 (2003)
21. A.G. Revesz, K.H. Zaininger: The Si-SiO₂ Interface, RCA Rev. **29**, 22 (1968)

3 Silicon: an Industrial Adventure

E.F. Krimmel

3.1 Introduction

The influence of industrial mass production of high-grade, single-crystal silicon on the development of our civilization was and is unique in the whole history of mankind: computers, communication systems, sensors, medical equipments, photovoltaics, satellites, space shuttles, etc. It could not have been predicted at all. Industrial companies worldwide, such as Bell Laboratories and Texas Instruments in the USA, and Philips, Siemens and Wacker in Europe, recognized their chance. However, the companies active in this field have also developed and changed. Today, a big portion of the relevant silicon is produced in Japan; the other part is shared between companies in the USA and Wacker in Europe. A brief, concise summary from the beginning up to the newest developments can be found in [1] and references cited therein.

3.2 The Principal Processes

The Siemens C-process (see Chap. 2), developed in the 1950s to produce high-grade single-crystal silicon, and the purification of the silicon by multiple zone refining, the trichlorosilane float-zone (FZ) technique, were key techniques and were adopted by Wacker. The competing technique forged in the USA was the Czochralski (CZ) pulling process. The FZ wafers cut from silicon rods were initially used mainly for power devices, and the CZ wafers were used for integrated circuits designed by Texas Instruments and Fairchild. The initial diameters of the wafers were 3 cm to 5 cm; today a diameter of 30 cm is already a standard dimension, and we are looking for diameters larger than 30 cm, say 40 cm, in order to increase the yield. Fortunately, the risky pioneering work went ahead in spite of “far-sighted experts” who were convinced that large-diameter wafers and 1 Mbit memories would never be needed and therefore never be produced.

Large-diameter wafers may exhibit a warping effect, due to high-temperature processes for instance. Thus whole-wafer contact printing to produce structures had to be replaced by projection printing and later by step-and-repeat single-chip printing techniques. The resolution was increased by applying light with ultrashort wavelengths, even in the UV or soft X-ray

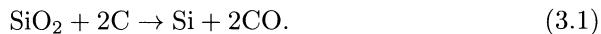
region, or by switching to electron or ion beam exposure with de Broglie wavelengths in the subatomic range.

Initially it was a dogma either to utilize the extremely pure FZ material or to utilize the oxygen-containing CZ material. However, it soon became clear that the CZ material was mechanically more stable than the FZ material because of dislocation pinning. This concerned, for instance, ion implantation, with its inherent problem of annealing radiation damage. The fracture of a wafer can also be reduced by simply rounding the wafer edge. Further, the oxygen in a wafer contributes to gettering deleterious impurities. The FZ-CZ dogma is nowadays smiled at, and only the proper quality of the silicon wafer is considered to be crucial for any particular application.

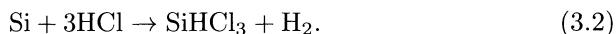
In order to reduce the costs of producing wafers, it has been proposed that wafers should not be made anymore of expensive high-quality single-crystal silicon but of a more economical form of silicon. The really needed thin, high-quality layer which incorporates the devices is then prepared by high-quality epitaxial deposition processes.

3.3 Some Details of Silicon Production

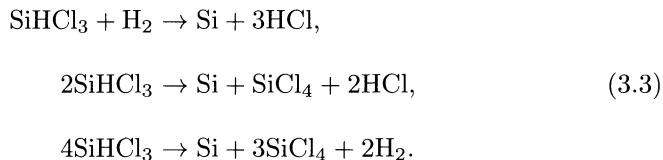
SiO_2 in the form of quartz sand is the starting material for producing electronic-grade silicon. First of all, this sand is reduced with carbon (coke) by an electro-thermal reaction at 2100 K to crude elemental silicon of a purity of approximately 98% in the presence of Fe to prevent the formation of SiC :



This silicon, in turn, is converted to high-purity trichlorosilane:



The trichlorosilane serves as the parent substance for forming ultrapure polycrystalline silicon on slim silicon rods by chemical vapour deposition (CVD) at approximately 1400 K, via the simultaneous reactions



Single-crystal silicon is then prepared applying, for example the crucible pulling (CZ) process by melting the silicon in a quartz crucible at 1415°C, adding dopants to the melt to adjust the resistivity, and pulling a crystal out of the melt using a single-crystal Si seed crystal forming a so-called neck so as to solidify the melt, with rotation, into a single-crystal silicon rod. The

diameter of the rod increases with decreasing pulling speed. This oxygen-containing silicon is mainly used to produce ICs. The float-zone pulling process is mainly used to prepare high-purity single-crystal silicon to be used to fabricate discrete devices such as power devices. This silicon is doped during the FZ process by exposing it to dopants in the gas atmosphere. Hyperpure silicon wafers are uniformly doped by neutron transmutation.

The rods are sliced into wafers by inner-diameter saw blades or multiwire slicing, the edges of the wafers are rounded, and the surface is lapped, etched to remove surface damage, chemically polished in a complex three-stage process, and possibly covered with a thin high-resistivity epitaxial layer.

Wacker, for instance, supplies wafers (up to 30 cm diameter) to the semiconductor industry today. This progress requires new equipment and new material characterization. The production of 45 cm wafers is already under discussion. The choice of CZ silicon is due in part to the mechanical properties. The weight of such a 45 cm rod may rise to approximately 400 kg, requiring larger-diameter necks, say of 1 cm to 2 cm. Hence, the complete technology must be subject to revision. Note that these wafers must be free of defects even on an atomic scale, apart from those which are present owing to the thermodynamic equilibrium conditions. Thus a rigorous control by defect analysis is required.

The industrial use of silicon, however, is not limited to microelectronics: silicon is used also in polycrystalline form for photovoltaics, chemical sensors, biosensors, physical sensors such as pressure sensors, flow sensors, mechanical sensors, temperature sensors, IR sensors, radiation detectors, Michelson interferometers, etc.

Note that amorphous silicon and even more, porous silicon are materials which are in development and show unexpected effects, which perhaps are not yet completely understood and are still under controversial. In particular, porous silicon became interesting to scientists because of properties such as light emission and its explosive property at low temperatures in the presence of hydrogen. It may be an attractive field for future industrial activities.

Acknowledgments

Relevant information from H. Silbernagel, E. Sirtl and V. Braetsch, all of Wacker, Burghausen, is gratefully acknowledged.

References

1. E.F. Haller: *The Semiconductor Age* (UC Berkeley and LBNL, USA 2002)

Part II

Polycrystalline Silicon

4 Polycrystalline Silicon Films for Electronic Devices

A. Slaoui, P. Siffert

4.1 Introduction

Polycrystalline silicon (poly-Si) plays an important role in the electronic devices of today. Several advantages are offered by poly-Si, which makes it one of the most important fundamental developments in the history of integrated circuits. First, from a structural point of view, poly-Si matches the mechanical properties of single-crystal Si, has good step coverage if deposited by CVD, has a high melting point, forms an adherent oxide, and is compatible with HF. From the device fabrication side, it absorbs and re-emits dopants, it neutralizes heavy metals (gettering), it has a compatible work function for MOS devices, and it forms high-conductivity silicides.

These properties make poly-Si widely used as an electronically active layer in integrated circuits (ICs) [1] and, more recently, in large-area electronic device technologies such as active-matrix liquid crystal displays (AMLCDs) [1, 2], silicon-based solar cells [3], and transducers [4, 5].

For IC processing, poly-Si offers many advantages because it can be selectively deposited, permits formation of shallow, high-quality junctions, fills trenches, permits salicidation, and provides a low MOS threshold voltage [6]. These advantages have resulted in many device applications, including self-aligned MOS gates, source/drain contacts, multilayer interconnects, dielectric isolation, nonvolatile floating gates, and emitter diffusion sources to form shallow junctions.

Subsequent processing for polysilicon gates involves doping, etching, and oxidation. In some device structures a second polysilicon layer is deposited. This layer may be used as a contact material in small windows or as an interconnect between conducting features.

The interest in polycrystalline silicon devices is also driven by the rapid growth of large-area electronic systems, specifically, flat-panel active-matrix displays and two-dimensional imagers [7], usually based on standard hydrogenated amorphous Si (a-Si:H) material. Poly-Si can satisfy many of the requirements because of its higher carrier mobility (~ 100 versus $1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ for a-Si) and the availability of good p-type polysilicon TFTs, thin film transistors, enabling high-performance CMOS, complementary metal-oxide-semiconductor circuits.

On the other hand, the fabrication of solar cells based on thin, polycrystalline Si films on foreign substrates (glass, metal, or ceramics) appears to be one of the most attractive routes to realizing cheap and efficient photovoltaic devices [3, 8, 9]. The challenge is poly-Si material with controlled resistivity and high minority carrier lifetime.

For these two large-area applications, the requirements on the material structure are rather different. In the case of TFTs, poly-Si films with a homogeneous grain size distribution are required in order to obtain reproducible electrical properties of the transistors. For thin-film solar cells, large-area polycrystalline films are required with a grain size of several μm or even several times $10\text{ }\mu\text{m}$, depending on the cell concept. For solar cells, the position of the grains is not critical. Moreover, there is no defined requirement on the grain size distribution, provided only that grains below $1\text{ }\mu\text{m}$ in diameter are absent.

Finally, there is a growing interest in poly-Si-based microcircuit transducers [1, 10] such as pressure sensors, strain gauges, vapor sensors, and mechanical components. The main advantages are reduction in material consumption and compatibility with IC processing.

4.2 Classification of Polycrystalline Silicon Films

Polycrystalline silicon films are composed of independent crystalline grains which are bounded by interfaces with the substrate or with the adjacent grains. A schematic view is shown in Fig. 4.1. Their properties and characteristics are inherently those of the inside of the grains and those of the interfaces. Therefore the grain structure is of importance in the application of polycrystalline thin films as materials. Thus, the average grain size should be as large as possible if the properties of bulk crystals are essential. Some applications require shrinking the grain size so that it is smaller than the area of the thin film used. If properties specific to a certain grain size are required, the size distribution of the grains must be as narrow as possible at that size. If both a large grain size and spatial uniformity are needed at the

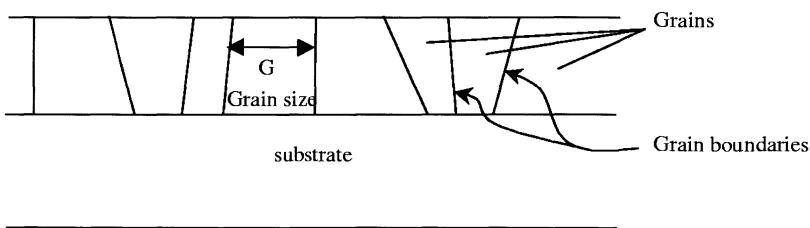


Fig. 4.1. Schematic cross section of a polycrystalline silicon film on a substrate

Table 4.1. Classification of polycrystalline Si films according to grain size

	nc-Si	pc-Si	pc-Si
Average grain size G	1 nm–50 nm	0.01 μm –1 μm	0.5 μm –1 mm
Growth temp. T_g ($^{\circ}\text{C}$)	$150 < T_g < 300^{\circ}\text{C}$	$250 < T_g < 500^{\circ}\text{C}$	$> 500^{\circ}\text{C}$
Electron mobility ($\text{cm}^2/\text{V s}$)	1–10	10–50	50–500
Hole mobility ($\text{cm}^2/\text{V s}$)	0.01–0.1	0.1–10	10–200
Minority carrier diffusion length	$< 0.1 \mu\text{m}$	$< 1 \mu\text{m}$	1–20 μm
Applications	NMOS, solar cells	CMOS, TFTs, solar cells	PMOS, CMOS, TFTs, sensors, solar cells

same time, one has to control the location of grains and their boundaries. In any case, the control of the grain structure is indispensable for achieving high-performance polycrystalline thin-film materials.

In most cases, the grain structure (size and distribution) in a polycrystalline silicon film and its subsequent electronic properties are strongly related to the processing temperature used to create the film and/or to the subsequent thermal treatment. This is clearly seen in Table 4.1, which presents a classification of crystalline Si films on a substrate according to grain size. “Polysilicon” or “poly-Si” is used as a generic word, whereas “nc-Si”, “pc-Si”, and “pc-Si” correspond to a range in grain size. The material required depends on the application.

The present review starts with a report on the different techniques used for the formation of polycrystalline Si thin films. These include direct (CVD, ...) as well as indirect (SPC, solid phase crystallization, metal-IC, metal-induced crystallization, laser annealing, ...) methods. A closer look at the structural properties of these Si films allows one to distinguish basic types of polycrystalline Si films with specific electrical properties. This overview discusses the specific limitations on performance associated with these properties.

4.3 Growth and Microcrystalline Structure of Poly-Si

Polycrystalline silicon has been grown by a variety of techniques, each suited to the needs of particular semiconductor device technologies [1].

Thin layers ($< 0.1 \mu\text{m}$) are generally grown by vacuum deposition techniques such as sputtering, evaporation, or molecular beam epitaxy (MBE). The athermal nature of the deposition processes allows one to use low temperatures (20 – 600°C). These techniques find application where thermally unsta-

ble substrates are used (e.g. solar cells, thick-film transistors, and hard mask layers).

Layers in the thickness range 10^{-2} – $10\text{ }\mu\text{m}$ are usually grown at temperatures between 150 and 850°C by low-pressure chemical vapor deposition (LPCVD). Nearly all the polycrystalline silicon layers incorporated into silicon integrated circuits are now deposited in this way, because of the uniformity, reproducibility, and conformal nature of the layer.

Thicker layers (1–100 μm) can be grown in atmospheric chemical vapor deposition systems over the temperature range 900 – 1300°C ; at this higher pressure, however, uniformity of deposition depends on gas phase diffusion, thus reducing the number of wafers processed and therefore the throughput. These layers find application as materials for solar cell applications.

Besides the above methods, a number of deposition techniques have been studied in which the thermal activation energy is supplemented by other energy sources, including plasmas (PECVD, plasma enhanced chemical vapor deposition, HWCVD, hot wire chemical vapor deposition, and ECRCVD, electron cyclotron resonance chemical vapor deposition) and laser or lamp irradiation.

In general, the grain size of the poly-Si increases with deposition temperature and with film thickness, increasing from less than 10 nm for thin evaporated films to greater than 1 mm for thick layers grown by LPCVD followed by zone-melting recrystallization (ZMR). The microstructure is, however, a complex function of the deposition conditions, and will be discussed in relation to the technique used to form the poly-Si film.

4.3.1 Poly-Si by CVD

Growth/Deposition by CVD

In the early days (1963–1970), poly-Si formation was investigated by physical vapor deposition (PVD) techniques. However, x-ray and ion damage, unwanted impurities, nonuniform step coverage, thickness variation, and low throughput prevented widespread use of these methods. Later (1970–1976), cold-wall, atmospheric-pressure CVD (APCVD) reactors were used for poly-Si deposition. The principle of chemical vapor deposition can be found elsewhere [11] and will not be described here. In the case of APCVD, the deposition is generally mass transport controlled, so wafers must be in a low-packing-density configuration to allow access of gases to the wafer surfaces. Such reactors gave quite high thickness variations and low gas efficiency. Controlled grain size was possible, depending on the used gas mixture and processing temperature used, to the detriment of the thickness uniformity. Even with these limitations, Si-gate PMOS and NMOS ICs became a major factor in the semiconductor device market in the early 1970s.

In 1976, the low-pressure CVD (LPCVD) process was introduced to deposit poly-Si layers. With LPCVD pressures in the range 1–100 Pa, the deposition is kinetically controlled and wafers can be standing up, with a high

packing density [12]. This led to lower cost than with APCVD (by more than 90%) and to improved uniformities ($\pm 1\text{--}2\%$), and provided the optimum grain structure for good post-deposition doping control. Although the pressures used in LPCVD systems are three to four orders of magnitude lower than in APCVD systems, deposition rates are typically only one order of magnitude lower at similar temperatures. This is because the precursor gas is diluted in APCVD systems, so the partial pressures of the precursor gas differ by only an order of magnitude or so between APCVD and LPCVD systems.

As a result, the LPCVD system has been the primary means of poly-Si deposition for ICs during the last 25 years, and this will continue for many years to come. The LPCVD reactors are selected depending on the device applications. Hot-wall, higher-temperature (850°C) LPCVD at higher pressures (0.5–5 Torr) and larger wafer spacing offers the interesting possibility of higher growth rates for selective poly-Si deposition. Conventional LPCVD ($\sim 600^\circ\text{C}$) in a horizontal or vertical diffusion furnace permits filling of the deep trenches or stacked capacitor cavities that are required for advanced CMOS memory devices. Additional capabilities include multilayer, sequential, in-situ deposition with special cleaning and etching steps.

Among the alternatives to LPCVD introduced in the 1980s, there is rapid thermal CVD (RTCVD), which uses an optically heated system [13]. RTCVD techniques include limited reaction processing (LRP), in which the reactant gases are present in the reactor and rapid heating/cooling of the wafer is used to start/stop the reaction [14]. A special quartz wafer carrier is used, which holds two layers of wafers and forms a volume separate from the rest of the reactor into which the precursor can be injected. This avoids deposition on the walls of the reactor and allows a high chemical yield. The drawback is that RTCVD techniques are usually limited to a single-wafer reaction chamber, which is an issue for high throughput. Lower temperatures ($600\text{--}650^\circ\text{C}$) provide a desirable grain size but an unacceptably low growth rate. Conversely, higher temperatures ($900\text{--}1000^\circ\text{C}$) provide higher productivity, but with an undesirable large grain size. Multiple steps (low-temperature nucleation with higher-temperature growth) offer some compromises [12].

The APCVD, LPCVD, and RTCVD systems allow the fabrication of high-quality poly-Si layers but at quite high temperatures ($> 600^\circ\text{C}$). However, there is an increasing need for depositing at low temperatures ($< 500^\circ\text{C}$) and at high deposition rate. In the last decade, growth rates have been considerably enhanced by supplying additional energy from plasmas or lasers, as will be seen below.

Deposition Rate of CVD Poly-Si

The transfer of silicon from gaseous-phase compounds (e.g. silane) to a solid phase incorporated into a growing polycrystalline silicon layer on a planar substrate has been extensively studied. Over a wide range of substrate temperatures and gas composition and pressure, the deposition rate is found not

to be limited by the transport of silicon from the gas to the substrate surface; the rate-limiting step is the rate of migration of SiH_4 molecules across the polysilicon surface and their self-nucleation and growth. It is this process which is thought to determine the measured activation energy for deposition.

If you look in the literature, the deposition rates of polycrystalline silicon obtained in real chemical vapor deposition systems are not usually directly comparable, because of the effects of contaminants and also because the published data is seldom comprehensive enough to enable comparison for exactly the same values of the important parameters. Key parameters are the partial pressure and nature of the silicon compound (e.g. SiH_4 or SiCl_4), the partial pressure of hydrogen, the doping gas and inert gas, and the total pressure in the system, as well as the substrate temperature.

In LPCVD furnace deposition systems, which typically operate in the range 600–650°C, with pure silane gas (pressures 100–400 mTorr) and deposition rates of 5–20 nm/min, gas transport is not the limiting factor, and it is because of this that such systems with high reproducibility are used. At such temperatures, the nature of the substrate has little influence on nucleation or growth: the native oxide (5–15 Å) always present on silicon substrates is stable, and so growth effectively takes place always on an amorphous substrate, whether silicon, silicon dioxide, or silicon nitride is the nominal underlayer.

In LPCVD systems, the deposition rate increases rapidly as the temperature increases. The activation energies are about 1.7 eV (40 kcal/mole), which is somewhat higher than the values observed for atmospheric-pressure deposition. The difference is caused by changes in the desorption of the hydrogen produced in the reaction and by differences in the roles of mass transport and homogeneous reactions. At temperatures higher than 650°C, gas phase reactions, which result in a rough, loosely adhering deposit, and precursor gas depletion, which causes poor uniformity, become significant. At temperatures much lower than 600°C, the deposition rate is too slow to be practical.

The gaseous silicon species used for CVD has a strong effect on the deposition rate. To illustrate, Fig. 4.2 plots the deposition rates of Si films on several substrates as a function of the deposition temperature at atmospheric pressure for trichlorosilane (TCS) [15] or dichlorosilane (DCS) [16] diluted in hydrogen. Data for thermally oxidized wafers (t- SiO_2) and monocrystalline wafers (sc-Si) are also included. Different regimes which are characteristic of deposition of silicon using thermal CVD [11] are observed. At low temperatures, the deposition rate is surface-reaction-controlled and it follows an exponential law with temperature: $r \propto \exp(-E_a/kT)$, where $E_a = 1.4\text{--}1.7\text{ eV}$ is the activation energy for the overall reaction. At high temperatures, the temperature dependence is much smaller. In this regime, the gas-phase diffusion of reactants to the surface is the limiting factor for the growth rate. This regime is referred as the diffusion-controlled reaction regime.

For Si homoepitaxy, high deposition rates of up to 6 $\mu\text{m}/\text{min}$ and 0.6 $\mu\text{m}/\text{min}$ in TCS- H_2 and DCS- H_2 , respectively, can be obtained, which can be ascribed to a more rapid nucleation and binding of the arriving silicon

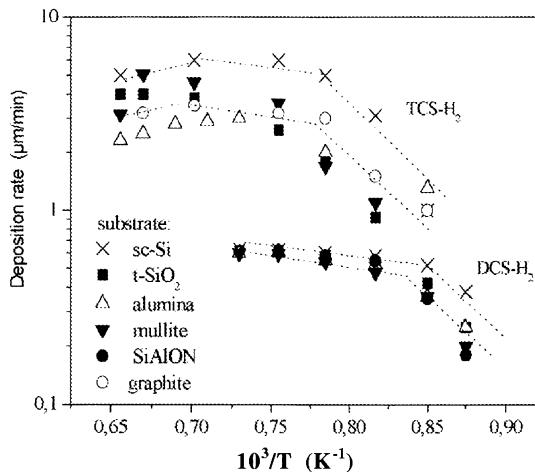


Fig. 4.2. Deposition rate of Si versus temperature on various foreign substrates. Depositions were carried out at atmospheric pressure with 15% TCS in H_2 [15] or 1.2% DCS in H_2 [16]

species to the Si atoms of the substrate. On the other hand, the deposition rate of poly-Si on t-SiO₂ and ceramic substrates in both regimes shows lower values when compared with deposition on sc-Si. Diffusion of reactive species on the substrate surface on one hand and the nucleation step on the other hand explain the difference in the kinetics of the deposition process between the various substrates.

Deposition rates from DCS are lower than from TCS. Indeed, the HCl interferes with the nucleation process and also reduces the surface mobility of the adsorbed silicon species. This is corroborated by the data in Table 4.2, which summarizes the growth rates from silane, dichlorosilane, trichlorosilane, and silicon tetrachloride measured over the temperature range 650–1200°C. Depositions were carried out at atmospheric pressure with 0.1% gaseous Si source in H_2 . In the surface-limited regime, all the activation energies controlling the temperature dependence were the same (1.45 eV), while

Table 4.2. Temperature range and corresponding deposition rate for different precursor silicon gases

Gaseous Si source	Temperature range (°C)	Activation energy (eV)	Preexponential factor (μm/min)
SiH_4	630–850	1.45	1.1×10^6
SiH_2Cl_2	800–950	1.45	3.0×10^5
SiHCl_3	800–1000	1.45	1.3×10^5
SiCl_4	850–1100	1.45	2.9×10^4

the preexponential factors decreased as the chlorine content of the species increased. At very high deposition temperatures, the formation of gaseous HCl facilitates the etching of the Si surface and therefore competes with the Si deposition.

Growth rates can be increased by increasing the silicon content of the gas molecule (e.g. Si_2H_6 versus SiH_4). Polysilicon can be doped during deposition by adding phosphine, arsine, or diborane to the reactants [17]. The dopant affects the deposition rate. For example, adding diborane causes a large increase in the deposition rate. Similar effects have been observed for deposition at atmospheric pressure. However, the thickness uniformity across a single wafer degrades when dopants are added. In fact, the in-situ doping method has not really found favor in horizontal tube reactors because of technical problems. Uniformity has been achieved by using an insert to control the flow of reactant gases around the samples.

Other species are known to act on the deposition rate (e.g. carbon and oxygen) and it is likely that these and other trace contaminants are responsible for the different values of the prefactor above which can be derived from the deposition rates published by different workers. The stability of the nuclei and the reproducibility of their occurrence are both worse at higher temperatures ($> 700^\circ\text{C}$) and in the presence of HCl.

Microstructure and Grain Size of CVD Poly-Si

The aim of direct deposition of polycrystalline silicon layers on nonsilicon substrates to obtain grains with a controlled size and distribution. To realize this, it is necessary to understand the laws governing silicon growth on amorphous substrates (oxidized silicon, glass, ceramics, . . .), where heterogeneous nucleation and grain growth are quite complex processes and in which many different mechanisms can play a crucial role. In the case of CVD, the final grain size is determined on one hand by the early-phase deposition, i.e. the nucleation phase, and on the other hand by competitive grain growth. During the nucleation phase, nuclei are formed and start capturing free atoms on the substrate surface; while these existing grains grow, new ones may be formed in the spaces between them. After coalescence, however, grains grow further epitaxially, continuing the underlying crystalline structure throughout the layer. The deposited film forms an amorphous structure if the arrival rate J of the silicon atoms is greater than the surface diffusion rate D , so that pairs of atoms form individual nuclei rather than migrating to larger nuclei. This happens up to a critical temperature T_c characteristic of each deposition system and deposition rate. T_c increases with deposition rate. Hydrogenated surfaces are reported to reduce crystallite nucleation and therefore to increase T_c [18]. The presence of other chemical species during deposition also affects T_c . At deposition temperatures above T_c , the structure could be partially or totally polycrystalline. In LPCVD systems, the films are fully crystalline at around 580 – 620°C and have a columnar structure, and the Si crystallites

have different crystal orientations [19]. They form grain boundaries when they reach each other. The relationship between N_x (the number of nuclei just before coalescence) and S_g (the grain size in the final layer) can be roughly approximated by $N_x = S_g^{-2}$. For 10 μm grains, for instance, the nucleus density to be obtained is 10^6 cm^{-2} . Thus it is the balance between nucleation and growth before coalescence which should be controlled to obtain a layer with the required average size. It is obvious that the density of nuclei and therefore the final grain size depend strongly on the substrate surface morphology and composition.

On the other hand, the continual increase in grain size S_g as the film thickness W increases has been simplified [20] to the relationship $S_g = 0.25 \cdot W$. The grain size is typically between 0.03 and 0.3 μm . The simple relationship is only a very rough guide, partly because it fails to take into account the effect of ambient pressure and temperature on grain size, but also because the actual grain structure is bimodal, containing equiaxed as well as columnar grains. The grain size after crystallization depends also on the dopant concentration. Polysilicon doped with a high concentration of phosphorus and heated between 900 and 1000°C for 20 min has an average grain size of 1 μm .

The crystallization temperature and, consequently, the final grain size are also reported to be affected by the presence of contamination [21]. Oxygen, nitrogen, or carbon impurities stabilize the amorphous structure to temperatures above 1000°C, and arsenic stabilizes the columnar structure to 900°C. Oxygen contamination is reduced as the deposition rate increases and there is an associated increase of grain size in polysilicon layers thicker than a few microns [22]. A similar effect is reported for carbon incorporation in CVD polysilicon growth. Occasional very large grains have been attributed to enhancement of growth by localized contamination. These results suggest that some of the differences in the observed grain size may be due to the different levels of purity of the deposition gases used in each investigation and the different surface purities of the starting substrates.

At high growth temperatures and pressures, Si nucleation is an adsorption process which is determined by the interaction between the impinging silicon atoms or a silicon atom cluster and the substrate surface [23]. It is critically influenced by the nature of the substrate because of the energy barrier for adsorption, which is different from one substrate to another [24]. Much larger grains can be obtained under these conditions. Table 4.3 compares the grain size and growth rate of different CVD systems.

It is usual that polysilicon films deposited at 600–650°C have a {110} preferred orientation [25]. At higher deposition temperatures the {100} orientation predominates, but the structure contains significant contributions from other orientations, such as {110}, {111}, {311}, and {331} [26]. Dopants and impurities, as well as temperature, also influence the preferred orientation.

Table 4.3. Grain size and growth rate for different CVD systems and gases used

Process	APCVD	LPCVD	PECVD	HWCVD	ECRCVD
T (°C)	800–1300	550–850	150–300	150–450	100–400
Precursor	SiH ₄ , Si gas	SiH ₄ , DCS, TCS, DCS, TCS	SiH ₄ , Si ₂ H ₆	SiH ₄ , Si ₂ H ₆	SiH ₄ , Si ₂ H ₆
Substrate	t-SiO ₂ , graphite, ceramics	Glass, t-SiO ₂	Glass, t-SiO ₂	Glass, metal foil	Glass, metal foil, plastic
Growth rate ($\mu\text{m}/\text{min}$)	1–10				
(nm/min)		5–20	5–20	50–200	5–50
S_g (μm)	1–50 μm	0.05–1 μm	0.05–0.1 μm	0.05–1 μm	0.05–0.1 μm
Applications	CMOS, solar cells	CMOS, solar cells	Solar cells	Solar cells	Solar cells
References	[15–17]	[38, 39]	[27–31]	[32–35]	[36, 37]

Alternatives to Thermal CVD

Although LPCVD systems allow the fabrication of high-quality poly-Si layers, at low temperatures (< 600°C) the deposition rate is generally too low. The deposition rate can be scaled up by plasma activation of the precursor gas. With plasma-enhanced CVD (PECVD), some of the energy required to break chemical bonds is provided by the plasma, so the temperature required to achieve a given growth rate can be lower [27, 28]. The high-energy electrons in the plasma collide with and dissociate gas molecules, which initiates the chemical reaction. In addition, the bombardment of the wafer surface by positive ions from the plasma can change the surface chemistry, resulting in different film structures and growth rates. RF glow discharges, which result in weakly ionized plasmas, are most commonly used for PECVD. Depending on the reactor configuration, the substrates are either within or downstream of the plasma. If the plasma is downstream of the plasma, improved control of the reaction chemistry can be achieved but it is often only possible to deposit on one substrate at a time. A disadvantage of PECVD is that the plasma can cause surface damage during deposition. When PECVD is performed at 200–300°C the result is relatively high-quality microcrystalline Si ($\mu\text{c-Si}$). The first report on this material was published in 1968 by Veprek and Marecek [29]. Overviews of the structural, optical, and electrical properties of $\mu\text{c-Si:H}$ material can be found in [30]. However, the material has a high hydrogen content (> 10% atomic) [31]. This is a drawback in the process, as bubbles appear in the film as a result of hydrogen evolution at high temperature and this leads to macroscopic defect creation in the bulk. This is why $\mu\text{c-Si:H}$ never gained much interest for application as a photovoltaically active material in solar cells or for ICs; it was, however, widely used for the doped layers of hy-

drogenated p-i-n solar cells, with hydrogenated amorphous silicon (a-Si:H) as the photoactive layer. A low hydrogen content is beneficial for suppressing spontaneous nucleation during deposition. Therefore temperatures of about 500°C are used [29].

Another plasma-assisted CVD technique is hot-wire CVD (HWCVD), where source gases such as SiH₄ and H₂ are pyrolytically decomposed on a filament catalyzer which is heated to about 1300–2000°C and located several centimeters from the surface of the substrate [32–35]. Gas-phase reactions and thin-film deposition then take place from the atomic and molecular precursors generated at the filament surface. The process takes place under high vacuum (~ 10 Pa). Doping can be readily achieved by adding B₂H₆ or PH₃ to the source gas. HWCVD is a simple and low-cost procedure. Single-step fabrication of poly-Si with reasonable deposition rates and good doping control is achievable. Large-area deposition is attainable by optimized superposition of precursors in a multiple-filament configuration.

One promising CVD method is electron cyclotron resonance CVD (ECRCVD) [36, 37]. This is a form of PECVD which uses ECR to produce the plasma. Cyclotron resonance occurs when the frequency of an alternating electric field matches the frequency of electrons orbiting the lines of force of a magnetic field. The plasma densities are higher with ECRCVD than with conventional RF PECVD. A specific advantage of ECRCVD is that it causes little surface damage, since the plasma source and substrate are well separated, and the operation pressure and plasma potential are low. In addition, substrate biasing allows separate variation of the plasma current and particle energy, there is the possibility of in-situ substrate pretreatment and layer post-treatment, there are no hot filaments or active electrodes, a higher proportion of the process gas is used, and there is the possibility of upscaling. A limitation of ECRCVD is the need for a very low pressure (0.1–1 Pa) and a high-intensity magnetic field, which increases the cost of the system.

4.3.2 Poly-Si by Crystallization of Amorphous Si

Formation of polycrystalline silicon films by deposition of a precursor film followed by crystallization has been extensively investigated over the last 25 years. The main driving force is the use of such films in macroelectronic devices such as large-area displays and solar cells. The thermal budget for processing poly-Si needs to be very low to enable the use of cheap and/or flexible substrates such as glass [38, 39] and organic polymers (plastics) [40]. Once amorphous silicon or fine-grained polysilicon has been deposited, there are several improvements in quality which may be made by further crystallization. The main high-temperature recrystallization technique is zone melting recrystallization (ZMR) [41]. Low-temperature crystallization techniques include solid-phase crystallization (SPC), metal-induced crystallization (MIC), and laser crystallization (LIC). Although crystallization adds another step to the process, it allows the use of a lower-cost, lower-quality initial silicon

deposition. These poly-Si layers may be used as either a seed layer or a base material for the device.

The most extensively studied method to obtain poly-Si is the SPC [42] of amorphous silicon deposited by PECVD, sputtering, or simple evaporation. The advantages of using SPC of a-Si are that it is simple and cost-effective, requires a low process temperature, produces a relatively high-quality active layer, is easy to scale up, and allows the possibility of in-situ phosphorus doping. SPC may be performed at temperatures above about 500°C, and produces large grain sizes. However, throughput is low. The development of the grain distribution during SPC of Si is described in [43].

For all growth temperatures during SPC, there is generally a log-normal distribution of grain sizes [44], which inevitably results in a large proportion of unfavorably small grains: the average grain size is a factor 3–5 smaller than the maximum grain size. This is especially important for solar cell applications, since the open-circuit voltage decreases significantly if even a small proportion of the grains have a diffusion length which is small compared with the grains in the cell.

Pulsed rapid thermal processing (PRTP), is a very fast method of SPC [45], as is metal-induced crystallization. It has been reported that the SPC temperature of a-Si can be lowered by the addition of some metals, such as Al, Ni, or Pd [46]. Metals lead to a lower thermal annealing temperature for the crystallization of a-Si by forming a metal silicide, by eutectic-alloy formation, by migration of metal silicide, or by some combination of these. Originally, Liu and Fonash [47] introduced a thin Pd film deposited on predeposited a-Si, which reduces the crystallization temperature, but the crystalline quality was not very satisfactory for device applications. Yoon et al. [48] reported on the crystallization of a-Si by annealing at $\sim 500^\circ\text{C}$ for 20 h using an ultrathin Ni layer on it. The Ni atoms on a-Si, after thermal annealing, form octahedral NiSi_2 precipitates in the a-Si matrix. The NiSi_2 precipitates are formed at the initial stage of thermal annealing and act as sites for crystallization. Needle-like Si crystallites are formed by silicide-mediated crystallization (SMC) of a-Si and the migration of NiSi_2 precipitates through the a-Si network. A Ni-SMC poly-Si TFT using a Ni solution exhibited a field-effect mobility of about $105 \text{ cm}^2/\text{Vs}$. More recently, nickel-nucleated lateral solid-phase epitaxy was achieved by nickel particles applied by means of a nickel colloidal “ink” [49]. The amorphous silicon layer crystallized fully before the onset of random nucleation, with each nickel particle seeding one grain, achieving grain sizes $> 100 \mu\text{m}$. Within each grain, however, there were many low-angle subgrain boundaries that arose from the needle-like crystal growth.

In spite of the low-temperature crystallization of a-Si, this method suffers from drawbacks such as metal contamination of the crystallized Si matrix and the long time for crystallization. To lower the crystallization temperature to 400°C and/or to shorten the duration of crystallization, an electrical bias has been applied during the Ni-SMC of a-Si [50]. The crystallization

speed increases with increasing electric field strength, which results from the accelerated migration velocity of the NiSi_2 precipitates in an electric field. In the presence of an electric field, the a-Si has been fully crystallized at 400°C for 30 min or 500°C for 10 min. A poly-Si TFT made using Ni field-enhanced SMC poly-Si exhibited a field-effect mobility up to $120\text{ cm}^2/\text{V s}$.

To avoid contamination, aluminum has been preferred. The aluminum-induced crystallization (AIC) method has been investigated to crystallize amorphous silicon layers deposited on glass [51] and ceramics [52]. Layers are made by thermally evaporating Al to a thickness of $0.5\text{ }\mu\text{m}$ and depositing a $0.5\text{ }\mu\text{m}$ layer of silicon by DC magnetron sputtering or PECVD. AIC is then done in a nitrogen gas ambient, with the result that the aluminium and silicon interchange as the silicon crystallizes, so that the structure is then glass/ceramic, poly-Si, Al+Si. At temperatures below 577°C , the silicon layer has a uniform thickness and forms separate crystals. Grains $10\text{--}20\text{ }\mu\text{m}$ in diameter are formed at an AIC temperature of 480°C . The grain size and speed of crystallization are also related to the average grain size of the Al [53].

Another extensively studied method for creating good-quality poly-Si on low-temperature substrates is excimer laser processing of amorphous silicon [54]. This is due to the fact that the excimer's short wavelength, high intensity, and narrow temporal pulse width ensure that thin a-Si films are melted and solidified rapidly, producing high-quality poly-Si before thermal damage of the substrate occurs. The rapid temperature changes achievable with pulsed laser illumination allow nucleation to occur near the melting point of silicon without melting a glass substrate. Since the nucleation rate is low near the melting point of silicon, grain sizes obtained by laser crystallization tend to be large.

The most established setup to produce large grains is the technique called sequential lateral solidification (SLS) [55]. The amorphous-Si layer is irradiated by an excimer laser through a patterned mask. After the exposed area has melted, the film is moved relative to the mask and another laser pulse is used to irradiate the silicon. This process has several remarkable features. Firstly, the substrate is at room temperature. Secondly, the process may lead to production of large single-crystal Si regions on glass. Thirdly, the log-normal distribution of grain sizes and the random grain boundary positions are avoided. Grain lengths of $200\text{ }\mu\text{m}$ have been produced thus. At present, the grain boundaries formed are essentially parallel. However, since the less energetically favorable grains eventually die out, the grain width increases slightly as the distance from the crystallization origin increases [55].

A combination of two techniques, lateral epitaxial growth and lateral explosive crystallization, has also been used to crystallize a thin layer of amorphous silicon deposited on glass [56]. The crystallized layer is used as a seed layer for an epitaxial thickening process. The principle of both techniques is to heat the silicon film while avoiding direct heating of the substrate. The lateral epitaxial solidification involves a $700\text{ }\mu\text{s}$ pulse from an Ar^+ laser. The melt crystallizes epitaxially after several hundred microseconds of exposure.

Crystallization begins at the outer rim of the melt pool, and the surface is corrugated and has few defects. Crystal grains around 100 μm long have been achieved.

Despite yielding excellent thin-film transistors [57], the width of these grains of several μm is too small for thin-film solar cells [58].

4.3.3 Chemistry of Grain Boundaries in CVD Poly-Si

In the 1980s, a number of reviews of the structure and properties of grain boundaries (GBs) in silicon were published, as well as several conference proceedings devoted to the same topic [59]. This was stimulated by the wide use made of this material as an interconnect in integrated-circuit metallization systems, and of rather larger-grained material in silicon solar cells. The hope is to find a link between the GB structure/chemistry and the electronic properties of these materials.

The current understanding of the structure of grain boundaries in polycrystalline silicon is that the core of some highly symmetric boundaries can be considered as a two-dimensional array of characteristic structural units. Because some of these units are highly distorted from the six-membered ring unit characteristic of the perfect silicon lattice, they are expected to offer preferential segregation sites for impurity atoms present in the grain interiors. Segregation to grain boundaries can alter the electronic properties of polycrystalline semiconductor materials quite dramatically, and it has been suggested that the electrical activity of grain boundaries is wholly a result of impurity segregation [60].

One of the most widely studied impurity segregation phenomena in silicon grain boundaries is that of oxygen diffusion to the grain boundaries during annealing treatments at over 600°C [61]. The increasing concentration of oxygen at the grain boundaries as the annealing time or temperature is increased has been linked to a larger grain boundary potential barrier. This decreases the conductivity of the polycrystalline aggregate, and also increases the rate of recombination at the boundaries. It is thus expected that annealing any large-grained material intended for solar cell production will result in a strong degradation of device performance. Indeed, most of these materials contain a high oxygen concentration. It is still not clear whether the oxygen forms oxide precipitates at the GBs [62] or is present as individual segregated atoms along the boundary plane. The more severe the segregation, the more likely it is that precipitates will be formed.

Fluorine seems to segregate to silicon GBs as well, and has a behavior similar to that of oxygen [63]. Titanium and aluminum have also been shown to segregate to silicon GBs, and the segregated boundaries appear to have a higher potential barrier than “clean” boundaries. On the other hand, dopants such as phosphorus and arsenic segregate strongly to silicon GBs to a saturation level between 0.1 and 1 monolayer. This is an equilibrium segregation process. Interestingly, little evidence has yet been found for the segregation of boron to silicon GBs.

Although most segregation phenomena have been observed after high-temperature heat treatments for relatively long times, as-deposited CVD polysilicon has also shown dopant segregation [64]. It is strongly thought that segregation phenomena have a significant influence on the electrical properties of both large- and small-grained polysilicon material.

On the other hand, the passivation of GBs in silicon sheets with hydrogen has also been widely investigated following the identification of dangling bonds at boundaries [65]. It has been assumed that the hydrogen will saturate the dangling bonds and lower the electrical activity at the boundaries. Clear observations have been made of the segregation of hydrogen to GBs in silicon [66]. Hydrogenation treatment has been shown to decrease the potential barriers and the recombination rates at GBs, and to increase the efficiency of silicon solar cells. The replacement of oxygen in silicon GBs by hydrogen has also been observed, although the nature of the exchange reaction between the two elements has not yet been determined. Kazmerski [2] has suggested that hydrogen segregating to oxygen-rich GBs saturates the oxygen-induced dangling bonds by the formation of hydroxyl groups. This reaction has been widely tested for improving the performance of polycrystalline silicon solar cells, whatever the grain size may be [67]. However, the extrapolation of information about the structure of grain boundaries obtained from large-grained material to describe the electrical properties of CVD polysilicon films before and/or after hydrogenation must be treated with caution. In fine-grained material it is common to ignore the problem of identifying the structure of the GBs altogether and to think of the boundaries as all having the same potential barrier, which is quite wrong.

One important characteristic of GBs is their orientation. Owing to the (110) texture, many boundaries should be of [110] tilt type. Such GBs are electrically inactive, and they grow without broken bonds. This is the case for nano- and microcrystalline Si materials. These [110] tilt type boundaries become active if they contain either impurities (e.g. oxygen) or structural defects which disturb the pure tilt behavior either by lattice dislocations or by twist components. It seems that growing Si with a (100) surface texture is the key to obtaining GBs with a low density of crystallographic defects. It is not an easy task, but it is needed for obtaining good thin-film transistors and highly-efficient solar cells.

4.3.4 Doping of Poly-Si

The incorporation of dopants into polysilicon is important because of the key role of polysilicon layers in VLSI technology (as an interconnect, as a gate material, and as a diffusion source to the underlying substrate) and in solar cell technology. The doping process has been mainly carried out by diffusion, implantation, or the addition of dopant gases during deposition (in-situ doping) [21,68]. Usually, a good correlation is found between the resistivity of the diffused polysilicon and the dopant solubility. Diffusion of dopants is faster

in polysilicon than in single-crystal silicon, and lateral diffusion in a polysilicon film is faster than diffusion perpendicular to the surface. This increase is attributed to the presence of grain boundaries in the poly-Si, which provide high-diffusivity paths through the material. Measurements of the diffusivity of dopants within polysilicon must therefore take account of two parallel processes: diffusion at and along the grain boundaries, and diffusion within the bulk interior of the grains. This information is very important for applications which require knowledge of the diffusion both vertically and laterally in the polysilicon, and in the underlying substrate [69].

The reported measurements of diffusivities in polysilicon can be divided into two categories: those that apply a single effective diffusivity to describe the combined effects of the grain interior and the grain boundary; and those that analyze the two separate processes, providing values for the diffusivity at the GB and in the grain interior. Measurements in the later category generally require a 2D analysis of the dopant distribution around a single grain boundary, and so large-grain material (cast or laser-enhanced recrystallized CVD) is used. Effects such as the segregation of dopants between grains and GBs, and the motion of GBs during heat treatment have also been considered [70].

It is also possible to take advantage of preferential diffusion of dopants along grain boundaries in order to increase the effective minority carrier diffusion length in fine-grained poly-Si [71]. Moreover, the grain boundaries with a high diffusivity are also those with a high recombination velocity. Converting the grain boundaries into n-type material (in the case of phosphorus doping, for instance) therefore decreases the number of recombination centers where electrons in the base can recombine. Furthermore, deep phosphorus diffusion enhances the gettering effect. All these effects should have strong effects on the collection of the carriers and therefore on the cell efficiency [72].

From the device point of view, a comparison of the three doping processes shows that the major differences are a lower resistivity for diffusion, a lower dopant concentration for implantation, and a lower mobility for in-situ doping. Implantation and in-situ doping, however, offer the advantage of lower processing temperatures, which is often the dominant consideration in VLSI processing.

4.4 Electronic Properties of Poly-Si

Electronic transport in poly-Si is considerably different from transport in single-crystal Si. It depends on the microcrystalline structure (grain size and orientation), the doping level in the grains, the grain boundary chemistry, and the defect density. Electronic charges at the grain boundaries control the material quality, and finally the device performance. Pinning of the Fermi level at deep states around midgap is the reason for the high sensitivity of Si to defects.

A band diagram for a GB in Si, belonging the so-called symmetric depletion type, is schematically shown in Fig. 4.3 [73]. For a classification and

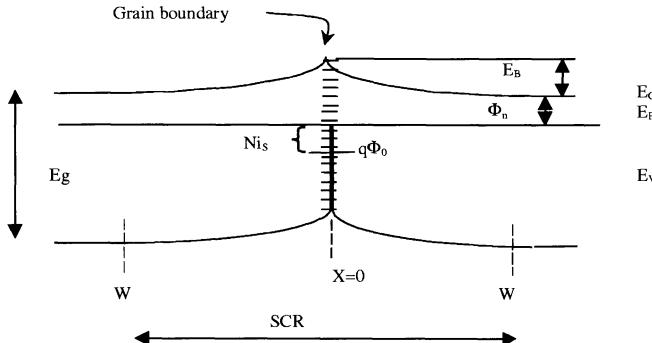


Fig. 4.3. Schematic illustration of band diagram for a GB in Si in the presence of a trap density N_{is} in n-type Si

a prediction of the electronic activity of GBs in different materials, see [74]. The barrier around the boundary hinders current transport of majority carriers. As a consequence, in a thin-film transistor the effective channel mobility is lower than for a single-crystalline material, whereas in poly-Si-based solar cells the interfacial charges attract minority carriers, enhance recombination, and thus lower the open-circuit voltage and efficiency [75].

The effects of the barrier height and defect density depend on the grain. For instance, recent theoretical and experimental investigations indicate that there is no significant band bending at grain boundaries of grains with diameters of 10 to 20 nm. The Debye screening length in Si is in the region of 100 nm and thus is much higher than the average grain size. The carrier transport in nanocrystalline Si is therefore dominated by trapping and recombination and not by the potential barriers at GBs. Thus the majority carrier mobility is of the order of a few cm^2/Vs and the minority carrier diffusion length is generally well below 1 μm . In this case, minority carrier devices such as solar cells are realized using pin structures in order to benefit from drift fields for carrier extraction.

In contrast, carrier transport in $\mu\text{-}$ or poly-Si films is governed by potential barriers at grain boundaries. The density of dangling bonds is quite sufficient to explain the observed potential-barrier heights at silicon grain boundaries, and models of grain boundary recombination and resistivity in poly-Si have been developed which give reasonable agreement with experimental measurements of the variation of these parameters with barrier height. A relationship between the resistivity and the barrier height, and the average grain size as well, has been established [76]:

$$\rho = \frac{(2\pi m^* k_B T)^{1/2}}{q^2 N_G S_g} \exp\left(\frac{E_B}{k_B T}\right),$$

where m^* is the “carrier” effective mass, and E_B is the barrier height.

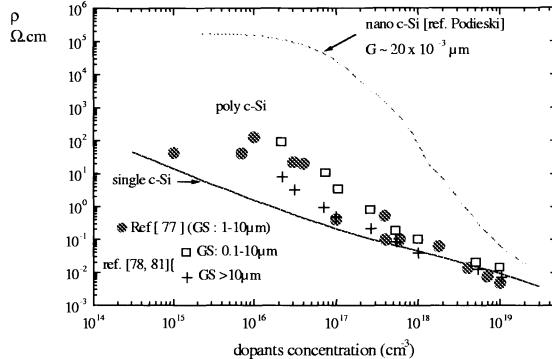


Fig. 4.4. Resistivity ρ as a function of acceptor concentration measured in polycrystalline silicon films of different grain sizes

As an illustration, Fig. 4.4 plots the resistivity of poly-Si layers with different grain sizes S_g versus the intragrain doping density N_G [77]. As expected, the resistivity of pc-Si is higher than that of single-crystal Si, especially at low doping levels. The smaller the grain size, the larger is the difference [78, 79]. This is attributed to trapping of carriers at the grain boundaries. Usually, the grain size S_g (μm), the trap density at GBs N_T (cm^{-2}), and the doping level N_G (cm^{-3}) are compared [80–82]:

- (i) For low doping levels ($N_G < N_T/G$), all carriers are trapped by the interface states at GBs (depleted grains), which leads to a quite low carrier concentration and therefore to a high resistivity.
- (ii) In the case where $N_G = N_{G*} = N_T/G$, the carrier concentrations in the grains and the traps at GBs are mutually neutralized. This corresponds to a sharp decrease in the mobility (Fig. 4.5). Increasing the grain size results in a shift of the mobility minimum to low doping levels.
- (iii) For high doping levels such as $N_G > N_T/G$, all traps at GBs are saturated, and the space charge region width is smaller than the grain size. A neutral zone appears in the grain in which the carrier concentration corresponds to the doping concentration, and becomes close to that of sc-Si as the doping concentration increases.

Thus, with typical values of $N_T \leq 10^{12} \text{ cm}^{-2}$ and average grain sizes $G \geq 1 \mu\text{m}$, the doping level is usually chosen $\geq 10^{16} \text{ cm}^{-3}$ to avoid partial or complete depletion of the grains.

Reduction of the potential barrier in as-grown poly-Si films through bulk passivation of the GB interface states by monatomic hydrogen has been widely studied. The efficiency of such treatment is not always straightforward. Fig. 4.6 plots typical curves of resistivity versus temperature for poly-Si layers before and after hydrogenation. These poly-Si films were deposited on silicon oxides in APCVD and LPCVD reactors (the average grain sizes were $4 \mu\text{m}$

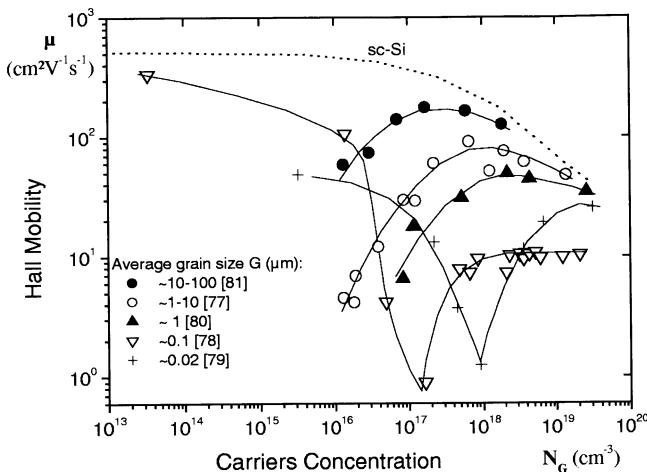


Fig. 4.5. Hall mobility versus carrier concentration for polycrystalline Si layers with different grain sizes. Data from [78–82]

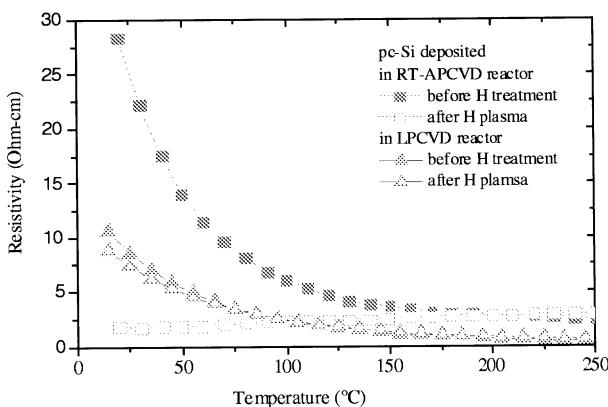


Fig. 4.6. Resistivity versus temperature, as measured before and after plasma hydrogenation (400°C, 1 h), for p-type polycrystalline silicon deposited on foreign substrates in RT-APCVD and LPCVD reactors

and 0.5 μm, respectively). For as-grown layers, the transport is dominated by the trap density, and the deduced (conductivity) activation energy E_B is in the range 180–200 meV. This corresponds to a density of positively charged interface states N_T at the grain boundaries of around $1 \times 10^{12} \text{ cm}^{-2}$. After hydrogenation treatment (400°C, 1 h), complete passivation of grain boundaries in the APCVD layer has been achieved, thus resulting in a significant resistivity reduction. The increase with temperature is characteristic of the mobility increase in silicon grains due to scattering by optical phonons. In the

small-grained (LPCVD) layer, however, only a small reduction ($\Delta \sim 20$ meV) is obtained after hydrogenation. In this case, the measured resistivity is governed by that at grain boundaries, with no conductivity contribution from the intragrain regions, which are still partly depleted.

In the case of solar cell applications, the electronic quality of the poly-Si is related to the lifetime of the free carriers generated in the bulk of the base layer. However, all CVD fine-grained polycrystalline silicon materials on foreign substrates show relatively short minority carrier lifetimes compared with multi- or monocrystalline silicon materials. Minority carrier diffusion lengths are usually in the range of only a few μm . The analysis of minority carrier properties of poly-Si films is intrinsically complicated by its inhomogeneous nature, which results from the broad grain size distribution and potential fluctuations within individual GBs. As a consequence, the open-circuit voltage of poly-Si thin-film cells is quite small (< 500 mV) [3, 8, 24]. In order to reduce the barrier height and thus charge carrier recombination, one has to reduce the doping of the absorber layer and/or the trap density at the grain boundary or increase the grain size. In the case of $\mu\text{-Si}$ -based solar cells, a pin-junction structure is employed instead. Furthermore, these materials contain several percent of hydrogen, and thus any remaining GB activity is passivated. Efficiencies of about 10.1% on glass have been reached. In large-grained Si ($\sim \text{mm}$) the spacing of grain boundaries is so large that potential barriers do not dominate recombination anymore, and a pn junction, with its higher open-circuit voltage potential, is usually used. Efficiencies in the range of 15–17% on high-temperature-resistant substrates [83] have been obtained.

4.5 Conclusion

The first application of a polysilicon gate in the MOS process, around 1970, was a crucial breakthrough for MOS technology, because it allowed the major advantages of polysilicon to be exploited. Since then, polycrystalline silicon has been used in the fabrication of all manner of devices, such as MOS, CMOS, and BiCMOC devices, bipolar transistors, displays, transducers, and solar cells, thanks to properties such as excellent compatibility with other materials used in silicon technology, temperature stability to over 1000°C, ease of doping and oxidation, and the ability to produce conformal edge coverage.

There are still some challenges in the formation of high-quality polycrystalline silicon films, such as:

- control of grain structure and doping
- enhanced conductivity
- integrated, multilayer processing
- high throughput, high growth rate, and large area (250 and 300 mm wafers).

Numerous new opportunities have still not been completely explored today, which will make poly-Si an important material for semiconductor devices well into the 21st century.

References

1. T. Kamins. In: *Polycrystalline Silicon for Integrated Circuit Applications*, Book serie vol. 45, ed. by Kluwer Academic Publishers (Kluwer Academic, Boston 1988)
2. p. 235 of [1]; see also articles in: *Flat Panel Display Materials I*, MRS Proc., vol. 345 (1994); *Flat Panel Display Materials II*, MRS Proc., vol. 424 (1996); *Advanced Materials and Devices for Large-Area Electronics*, MRS Proc., vol. 685 (2001)
3. J.H. Werner, R. Bergmann, R. Brendel. In: *Advances in Solid State Physics*, vol. 34, ed. by R. Helbig (Vieweg, Braunschweig 1994) p. 115
4. p. 246 of [1]
5. K. Ikeda. In: *Technical Digest of the 7th Sensor Symposium* (JISCST, Tokyo 1988) p. 193; see articles in: *Materials Science of Microelectrochemical Systems*, MRS Proc., vol. 657 (2001)
6. E.C. Douglas: Solid State Technol. **24**, 65 (1981)
7. J.G. Blake, M.C. King, J.D. Stevens, R. Young: Solid State Technol. **40(5)**, 15 (1999)
8. R.B. Bergmann: Appl. Phys. A **69**, 187 (1999)
9. A. Slaoui, J. Poortmans, T. Vermeulen, R. Monna, O. Evrard, K. Said, J. Nijs: J. Mater. Res. **13**, 2763 (1998)
10. R.T. Howe: Thin Solid Films **181**, 235 (1989)
11. S. Sivaram. In: *Chemical Vapor Deposition* (International Thomson, New York 1995)
12. M.L. Hammond. In: *Silicon Processing*, ed. by D.C. Gupta (ASTM, Boston 1983) p. 206
13. See multiple papers In: *Rapid Thermal Annealing/Chemical Vapor Deposition, and Integrated Processing*, MRS symposia, vols. 146 (1989), 224 (1992), 342 (1994), 387 (1995), 429 (1996)
14. J.F. Gibbons, C.M. Gronet, K.E. Williams: Appl. Phys. Lett. **47**, 721 (1985)
15. D. Angermeier: PhD dissertation, Louis Pasteur University, Strasbourg, France (1998)
16. A. Van Zutphen: PhD dissertation, Technische Universiteit Delft, The Netherlands (2001)
17. F.C. Everssteyn, B.H. Put: J. Electrochem. Soc. **120**, 106 (1973)
18. A.M. Beers, J. Bloem: Appl. Phys. Lett. **41**, 153 (1982)
19. R. Bisaro, P.N. Magarino, K. Zellama: J. Appl. Phys. **59(4)**, 1167 (1986) p.
20. C.P. Ho, J.D. Plummer, S.E. Hansen, R.W. Dutto: IEEE Trans. Electron Devices **30**, 1438 (1983)
21. T.I. Kamins: J. Electrochem. Soc. **127**, 833 (1980)
22. T.I. Kamins, M.M. Mandurah, K.C. Saraswat: J. Electrochem. Soc. **125**, 927 (1978)
23. J. Bloem: J. Cryst. Growth **50**, 581 (1980)

24. A. Slaoui, J. Poortmans, M. Maex. In: *Growth, Characterization, and Electronic Applications of Si-Based Thin Films*, ed. by R.B. Bergmann (Research Signpost, Trivandrum, India 2002) p. 147
25. R. Bisaro, P.N. Magarino, K. Zellama: *J. Appl. Phys.* **59**, 1167 (1986)
26. D. Angermeier, R. Monna, A. Slaoui, J.C. Muller: *J. Cryst. Growth* **191**, 386 (1998)
27. K. Fujimoto, F. Nakabeppu, Y. Sogawa, Y. Okayasu, K. Kumagai. In: *Proc. 23rd IEEE Photovolt. Spec. Conf.* (Electron Device Society, New York 1993) p. 83
28. O. Vetterl, P. Hapke, O. Kluth, A. Lambertz, S. Wieder, B. Rech, F. Finger, W. Wagner: *Solid State Phenomena* **67-68**, 101 (1999)
29. S. Veprek, V. Marecek: *Solid State Electron.* **11**, 683 (1998)
30. S. Veprek, M. Heintze, F.A. Sarott, M. Jurcik-Rajman, P. Willmott: *Mater. Res. Soc. Proc.* **118**, 3 (1988)
31. N. Wyrsch, P. Torres, M. Goerlitzer, E. Vallat, U. Kroll, A. Shah: *Solid State Phenomena* **67-68**, 89 (1999)
32. J.K. Rath, H. Meiling, R.E.I. Schropp: *Jpn. J. Appl. Phys.* **36**, 475 (1997)
33. H.N. Wanka, M.B. Schubert, A. Hierzenberger, V. Baumung. In: *14th European Photovolt. Solar Energy Conf.*, Barcelona (1997)
34. M. Ichikawa, J. Takeshita, T. Tsushima, A. Yamada, M. Konagai. In: *Technical Digest of the International PVSEC-11*, Sapporo (1999) p. 943
35. R.E.I. Schropp. In: *Thin Film Materials for Photovoltaics*, ed. by A. Slaoui, J. Poortmans, A. Jager-Waldau, C. Brabec, EMRS 2001 Spring Conference, *Thin Solid Films* **403-404**, 17 (2002)
36. P. Muller, I. Beckers, E. Conrad, L. Wistner, W. Fuhs. In: *25th Photovolt. Spec. Conf.*, (Electron Device Society, New York 1996) p. 673
37. K.E. Lee, W.H. Lee, S.S. Shin, C. Lee: *Jpn. J. Appl. Phys.* **35**, 258 (1996)
38. R. Rogel, K. Kission, T. Mohammed-Brahim, M. Sarret, O. Bonnaud, J.P. Kleider. In: *2nd World Conf. and Exhibition on Photovoltaic Solar Energy Conversion*, Vienna (1998) p. 1701
39. B. Caussat, J.P. Couderc, A. Figueras, A. Vander Lee, J. Durand, V. Paillard, E. Sheid, J.R. Morante: *Solid State Phenomena* **67-68**, 125 (1999)
40. S. Wagner, H. Gleskova, I.C. Cheng, M. Wu. In: *Growth, Characterization, and Electronic Applications of Si-Based Thin Films*, ed. by R.B. Bergmann (Research Signpost, Trivandrum, India 2002) p. 1
41. T. Ishihara: *ibid.*, p. 79
42. G.L. Olson, J.A. Roth: *Mater. Sci. Rep.* **3**, 1 (1988)
43. R.B. Bergmann, F.G. Shi, H.J. Queisser, J.Krinke: *Appl. Surf. Sci.* **123/124**, 376 (1998); H. Kumoni, F.G. Shi: *Phys. Rev. Lett.* **82**, 2717 (1999)
44. P. Kwizera, R. Reif: *Appl. Phys. Lett.* **41**, 379 (1982)
45. Y. Komem, I.W. Hall; *J. Appl. Phys.* **52(11)**, 6655 (1981)
46. S.Y. Yoon, K.H. Kim, C.O. Kim, J.Y. Oh, J. Jang: *J. Appl. Phys.* **82**, 5865 (1997)
47. H. Liu, S.J. Fonash: *Appl. Phys. Lett.* **62**, 2554 (1992)
48. S.Y. Yoon, K.H. Kim, J.Y. Oh, J. Jang: *Jpn. J. Appl. Phys.* **37**, 7193 (1998)
49. H.A. Atwater, C.M. Chen. In: *Growth, Characterization, and Electronic Applications of Si-Based Thin Films*, ed. by R.B. Bergmann (Research Signpost, Trivandrum, India 2002) p. 55
50. J. Jang, J.Y. Oh, S.Y. Yoon, K.H. Kim, C.O. Kim: *Nature* **395**, 481 (1998)

51. O. Nast, T. Puzzer, L.M. Koschier, A.B. Sproul, S.R. Wenham: *Appl. Phys. Lett.* **73**, 3214 (1998)
52. A. Slaoui, E. Pihan, M. Rusu. In: *17th European Photovoltaic Solar Energy Conf.*, Munich (October 2001) p. 1462
53. O. Nast, S. Brehme, S. Pritchard, A. Aberle, S.R. Wenham. In: *Technical Digest of the International PVSEC-11*, Sapporo (1999) p. 727
54. J.R. Köhler. In: *Growth, Characterization, and Electronic Applications of Si-Based Thin Films*, ed. by R.B. Bergmann (Research Signpost, Trivandrum, India 2002) p. 39
55. R.S. Sposili, J.S. Im: *Appl. Phys. Lett.* **69(19)**, 2864 (1996)
56. G. Andra, J. Bergmann, F. Falk, E. Ose, N.D. Sinh. In: *Technical Digest of the International PVSEC-11*, Sapporo (1999) p. 741
57. Y. Helen, R. Dassow, M. Nerding, K. Mourguès, F. Raoult, J.R. Köhler, T. Mohammed-Brahim: *Thin Solid Films* **383**, 143 (2001)
58. R.B. Bergmann, J.H. Werner. In: *Thin Film Materials for Photovoltaics*, ed. by A. Slaoui, J. Poortmans, A. Jager-Waldau, C. Brabec, EMRS 2001 Spring Conference (Elsevier, Amsterdam 2002) p. 162
59. S. McKernan, C.B. Carter: *Solid State Phenomena* **37-38**, 67 (1994)
60. C.R.M. Grovenor: *J. Phys. C* **18**, 4079 (1985)
61. L.L. Kazmerski, J. Vac: *Sci. Technol. A* **4**, 1638 (1986); G. Harbecke: *Polycrystalline Semiconductors* (Springer, Berlin 1985)
62. F. Battistella, A. Rocher, A. George: *Mater. Res. Soc. Symp. Proc.* **59** 347 (1986); B. Cunningham, H.P. Strunk, D.G. Ast: *J. de Physique* **43(C-1)**, 51 (1982)
63. D.S. Ginley: *Appl. Phys. Lett.* **39**, 624 (1881); F. Komninou, T. Karakostas, G.L. Bleris: *J. de Physique* **43(C-1)**, 9 (1982)
64. C.R.M. Grovenor: *Philos. Mag. A* **50**, 409 (1984)
65. N.M. Johnson, D.K. Biegelsen, M.D. Moyer: *Appl. Phys. Lett.* **40**, 882 (1982); C. Dianteill, A. Rocher: *J. de Physique* **43(C-1)**, 75 (1982)
66. L.L. Kazmerski, J.R. Dick: *J. Vac. Sci. Technol. A* **4**, 1120 (1984)
67. A.W. Weeber, H.H.C. de Moor, R.A. Steeman, W.C. Sinke, F.M. Schuurmans, P.-P. Michiels, L.A. Verhoef, P.F.A. Alkemade, E. Algra: *Solid State Phenomena* **37-38**, 355 (1994); N.H. Nickel, N.M. Johnson, W.B. Jackson: *Solid State Phenomena* **37-38**, 367 (1994)
68. G. Yaron: *Solid State Electron.* **22**, 1017 (1979)
69. B. Arab: *Sol. Energy Mater. Sol. Cells* **37**, 239 (1995)
70. A. O'Neill, C. Hill, J. King, C. Please: *J. Appl. Phys.* **64**, 167 (1987)
71. G. Beaucarne, S. Bourdais, A. Slaoui, J. Poortmans: *Proceedings of the 28th IEEE Photovoltaic Specialists Conference*, Anchorage (2000) p. 128
72. E. Christoffel, M. Rusu, A. Zerga, S. Bourdais, S. Noël, A. Slaoui: *Thin Solid Films* **403-404**, 258 (2002)
73. H.C. Card, E.S. Yang: *IEEE Trans. Electron Devices* **24(4)**, 397 (1977)
74. J.H. Werner, N.E. Christensen. In: *Polycrystalline Semiconductors II*, ed. by J.H. Werner, H.P. Strunk, Springer Proceedings in Physics, vol. 54 (Springer, Berlin 1991) p. 145
75. J.H. Werner, R. Dassow, T.J. Rinke, J.R. Köhler, R.B. Bergmann: *Thin Solid Films* **383**, 95 (2001)
76. C.H. Seager, T.G. Castner: *J. Appl. Phys.* **49(7)**, 3879 (1978)
77. S. Bourdais, G. Beaucarne, J. Poortmans, A. Slaoui: *Physica B, Condens. Matter* **273**, 544 (1999)

78. M.W.M. Graef, J. Bloem, L.J. Gilling, J.R. Monkowski, J.W.C Maes: Proc. 2nd EC PVSEC, Luxembourg (1979) p. 65
79. M.E. Cowher, T.O. Sedgick: J. Electrochem. Soc. **119**, 1565 (1972)
80. J.Y.W. Seto: J. Appl. Phys. **46**, 5247 (1975)
81. T.I. Kamins: J. Appl. Phys. **42**, 4357 (1971)
82. M.W.M. Graef, L.J. Gilling: J. Appl. Phys. **48**, 3937 (1977)
83. A.M. Barnett, R.B. Hall, J.A. Rand, C.L. Kendall, D.H. Ford: Sol. Energy Mater. **23**, 164 (1991)

5 Silicon for Photovoltaics

J.-C. Muller, P. Siffert

5.1 Introduction

Photovoltaics (PVs), which convert solar energy directly into electricity, are probably the most effective alternative energy source to conventional power supplies. The world demand for terrestrial applications of photovoltaic systems has increased, first for isolated areas and more recently in connection with grids in large PV roof programs, the most important cases being in Japan and Germany.

Many types of solar cell structures have been developed, based on various semiconductor materials, including II–VI or III–V compounds. Silicon was the first material used and is probably still the most appropriate candidate for the next 20 years of large-scale applications. The reasons for this success are the great availability, technical advance and environmental safety of silicon.

Solar cells produced from cast multicrystalline silicon (large-grain polycrystalline silicon, referred to as mc-Si) can be considered as the only present-day technology capable of achieving low-cost high-rate production without excessive loss of efficiency compared with single-crystal silicon (sc-Si). The global module-shipping market reached 530 MWp in 2002 [1] (Fig. 5.1), quite twice the market two years before, with 99% of the products still based on silicon and now more than 55% on cast mc-Si ingots [2] (see Fig. 5.2).

Polycrystalline silicon substrates contain grain boundaries and residual impurities in higher concentrations than in single-crystal silicon. For this reason, the photovoltaic efficiency was initially inferior to that obtained using conventional Czochralski (Cz) or float-zone (FZ) silicon.

However, 25 years of research on multicrystalline silicon have enriched the knowledge of the material and of its thermally activated defect reactions during cell processing, leading to a substantial improvement of the manufacturing techniques and to the capability to produce high-efficiency solar cells with this material. The difference between polycrystalline and single-crystal cells has since decreased continuously.

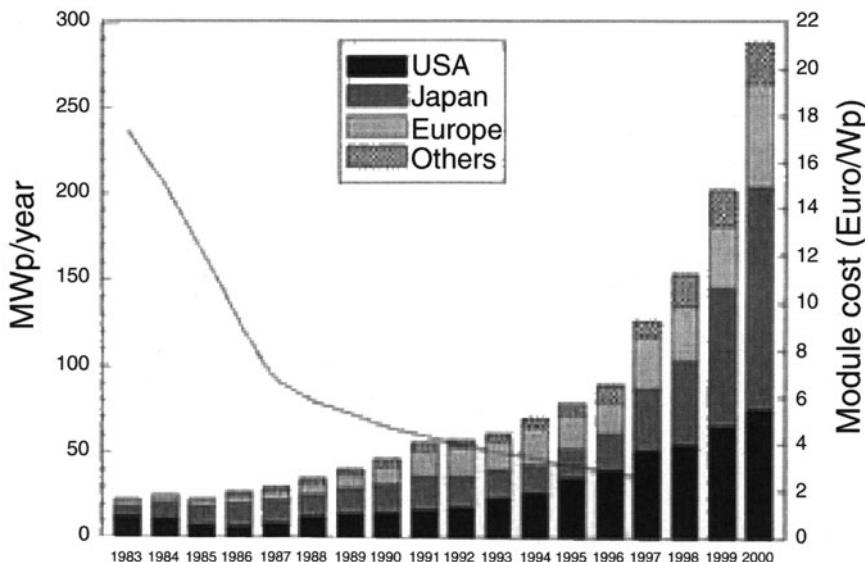


Fig. 5.1. Evolution of the cost per module and of the world module-shipping market. In 2001 the production reached 400 MWp, twice the 1999 market [1], and 530 MWp in 2002 [2] (also quite twice the year 2000 market), now with a shipping cost lower than 3 euros/Wp

5.2 Silicon Material for PVs

This part will be devoted to the material-growing technology. It covers the initially used Cz single crystals, cast multicrystalline ingots, silicon ribbons and the various thin-film deposition processes. It also addresses the emerging crucial problem of the preparation of cheap solar-grade feedstock material.

Although the ribbon and film processes can use electronic-grade silicon or pure SiH_4 gas as feedstock material, the cast-ingot technologies need cheap parent materials since about 60% of the material is lost during the slicing into 300 μm sheets. For these reasons, the PV industry has up to now used materials rejected by the microelectronics industry. However, since the PV market is growing faster than the microelectronic market, the cost of the parent material has multiplied by a factor of 3, now exceeding \$25 per kg. Consequently, the availability of silicon-grade feedstock material has become a crucial problem.

Today, simplified purification methods based on adapted chemical methods or derived from directional solidification processes that apply, for example, electron beam pulling methods [3] or plasma purification, are being investigated.

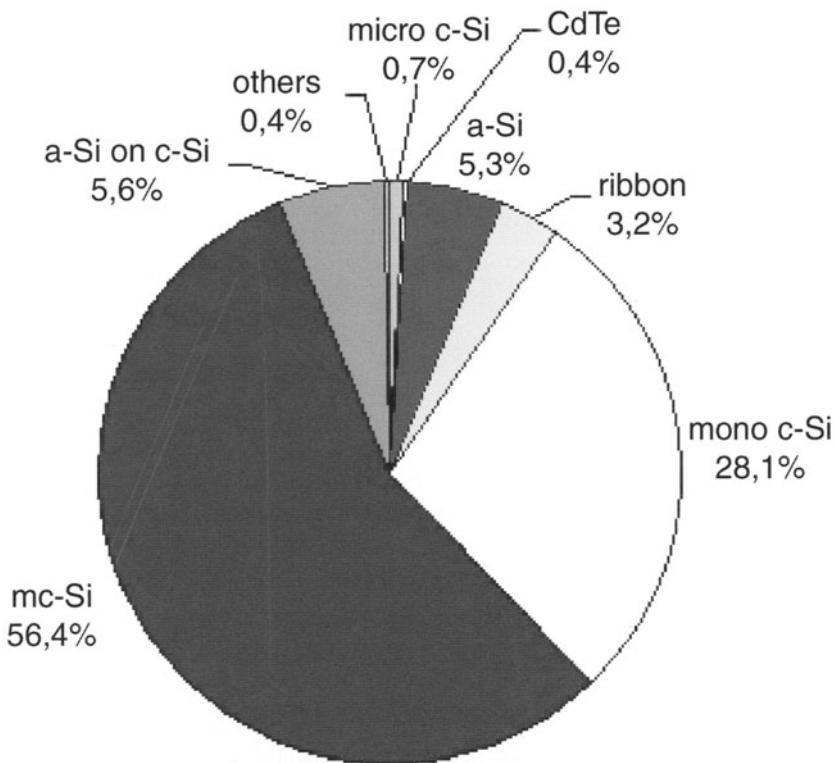


Fig. 5.2. Illustration of the percentage of the different type of materials in the total world production: The share of silicon in the total world product is still 99%, with more than 55% of it being cast mc-Si ingots [2]

5.2.1 History and State of the Art of Different Growing Processes

(a) Cast Ingots by Directional Solidification

In 1975, Wacker proposed a new method, called the Wacker ingot casting process (WICP) [4], to manufacture low-cost substrates for terrestrial solar cells, instead of the conventional Czochralski technology. Later on, many other casting processes were introduced by a large number of research groups and industrial companies throughout the world, such as Solarex (UCP) [5], Crystal-System (HEM) [6], CGE/Photowatt (Polix) [7], IBM (DS) [8], NEC/OTC (NMR) [9], Eurosolare [10], Crystalox [11] and Bayer [12].

These processes use quartz crucibles to melt silicon pieces from Cz ingots rejected by the microelectronics industry and subsequently apply a directional solidification. The solidification process is carried out with a controlled vertical temperature gradient of the silicon melt in the mould. This leads to a silicon block with vertical lines of crystalline silicon with a columnar structure

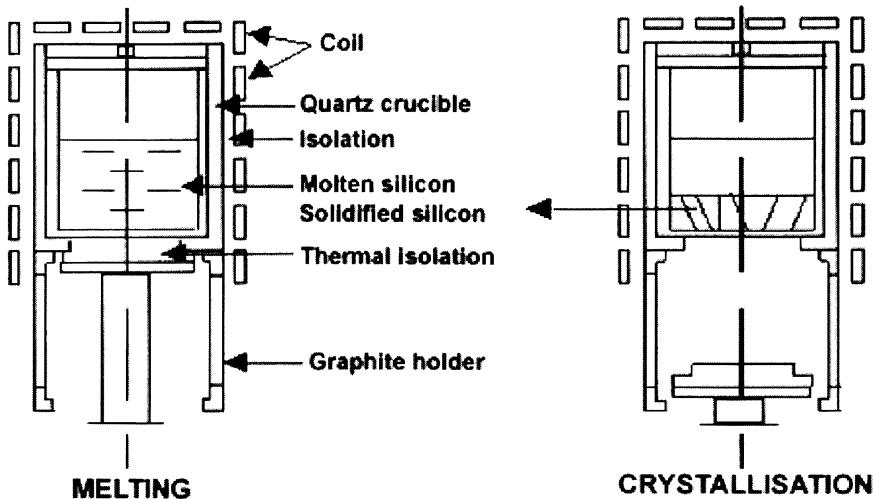


Fig. 5.3. Illustration of the solidification process carried out with a controlled vertical temperature gradient of the silicon melt in the mould to obtain a block of crystalline silicon with a columnar structure of vertical lines from the bottom to the top of the block, e.g. Polix pulling. From Photowatt International [13]

extending from the bottom to the top (see Fig. 5.3). The casting methods were rapidly proven to be cheaper than conventional Cz methods, but their main disadvantage was that they yielded materials containing stresses, impurities and grain boundaries.

Many methods have been investigated to improve the heat flux control during solidification, in order to yield interfaces as planar as possible and hence perfect columnar crystal growth. The manufacturers have gradually reduced the internal stresses and have been able to suppress the generation of cracks by a more careful control of the temperature gradient.

The manufacturers have also tried to avoid interaction of the molten silicon with the mould. Graphite moulds, which can be reused with an additional coating of the inner walls, induce significant contamination. They have been progressively replaced by single-use quartz moulds, as these allow manufacturers to reduce considerably the oxygen and carbon content of the material.

Moreover, in order to reduce the manufacturing costs, efforts to obtain larger blocks have been made. Ingots with cross-sectional areas larger than $50 \times 50 \text{ cm}^2$, a height of 18 or 28 cm, and 250 kg in weight are now currently available.

After contouring and removal of the bottom and top regions, which do not satisfy the conditions for use, the ingot is partitioned into 9, 16 or 25 elementary blocks having a base area between $10 \times 10 \text{ cm}^2$ and $15 \times 15 \text{ cm}^2$, using bandsaws. The final slicing into wafers with a thickness of about 300 μm is performed with multi-wire saws.

Both the progress in multi-wire sawing of very thin slices – Photowatt [13] was the first to cut at about 200 μm – and the further development of continuous casting are the main reasons for the success of cast ingots as compared with the ribbon technology.

(b) Electromagnetic Casting (EMC)

In 1981–1985, OTC [14] and SERI [15] developed a cold-crucible casting technique based on induction heating. Neither a crucible nor a mould is used in this method. The molten silicon is heated and confined electromagnetically by induction power in order to avoid impurity contamination from the furnace

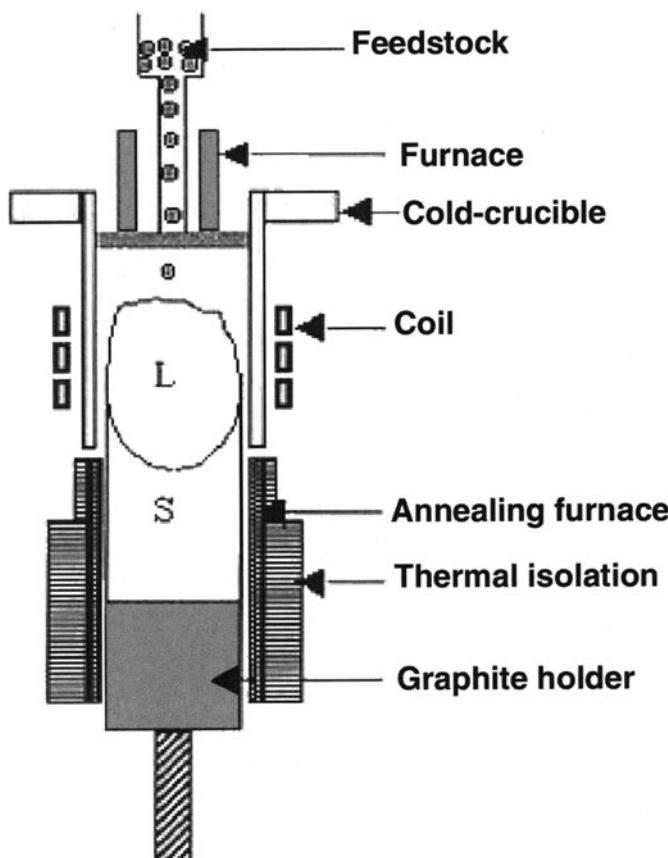


Fig. 5.4. Illustration of the cold-crucible casting technique based on induction heating. In this method, no mould is used. The molten silicon is heated and confined electromagnetically, for example in the Emix process, initially developed by EPM-Madlylam with Photowatt [17] and now used industrially by EMIX S.A. [18]

(Fig. 5.4). The EMC vertical continuous-pulling method was developed by Sumitomo SiTiX Corp. [16], formerly OTC, as well as by EPM-Madylam [17] together with Photowatt, and will be used by EMIX SA in the future [18]. Crystaloxy [19], on the other hand, has developed a horizontal-zone refining method for silicon applying induction heating.

Ingots of a cross-sectional area of approx. $35 \times 35 \text{ cm}^2$, 3 m long and with a weight of 850 kg can be obtained by these continuous-pulling techniques using plasma torch melting for better control of the solid/liquid interface and of contamination [20].

The cold-crucible casting technique, which reduces the energy consumption for crystal pulling by almost a factor of 4, constitutes a major opportunity for the future evolution of multicrystalline silicon, which is in constant competition with silicon ribbons and, more recently, with silicon thin films.

(c) Ribbon Silicon for PV

Silicon ribbon technology, another promising method that is in competition with both ingot casting and thin-film technologies, also produces multicrystalline material. The production of the first silicon ribbons was done simultaneously in 1975 by Westinghouse [21] with the dendritic WEB process called

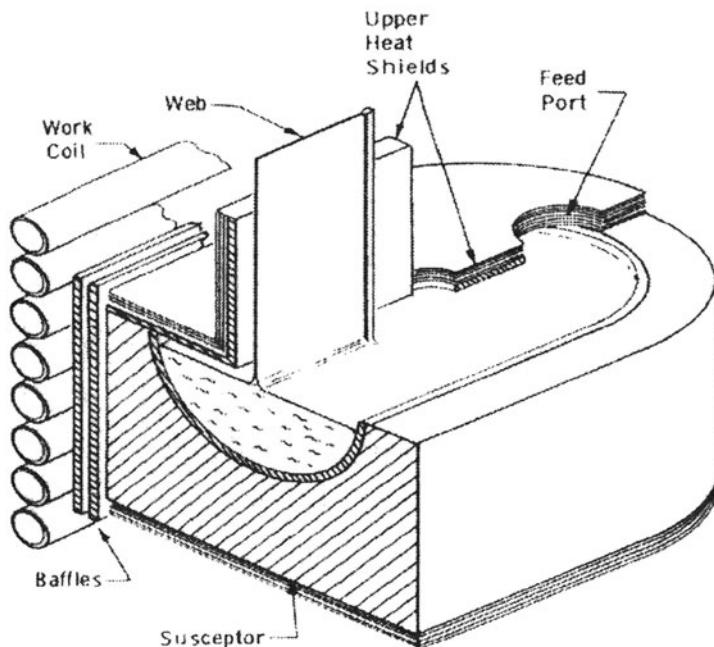


Fig. 5.5. Illustration of the WEB process for growing silicon sheets [21]

“WEB” ribbon process (Fig. 5.5), by ASE-Americas [22] (formerly Mobil-Tyco/Mobil-Solar) with the EFG, edge defined film fed grown [23] process using two graphite dies, and by Motorola [24] with the RTR, ribbon-to-ribbon crystal growth process using laser melting.

A process of two-sided film deposition on carbon sheets (RAD, ribbon against drop pulling process ribbon) was developed later by CGE [25] in France. The silicon grain structure obtained by all these processes shows more longitudinal grains compared with cast ingots. The crystallization quality depends strongly on the pulling velocity (generally around a few cm per minute) and on the pulling material, such as graphite or carbon sheets.

At present, the ribbon material represents 3.7% of the world market [26]. Research is going towards thinner substrates by developing 50 cm diameter hollow cylindrical tubes with a thickness of 75–80 μm [27]. In Europe, RWE-Schott-Solar, Germany, has recently started a $2 \times 30 \text{ MWp}$ factory based on the EFG technology [28].

5.2.2 The Thin-Film Deposition Processes

(a) Amorphous, Microcrystalline or Quasi-crystalline Silicon Thin Films for PVs

The thin-film technologies represent today about 9% of the world PV market, and this percentage is mainly accounted for by amorphous silicon. Figure 5.6 [29] reports all laboratory cell efficiency results reported for the var-

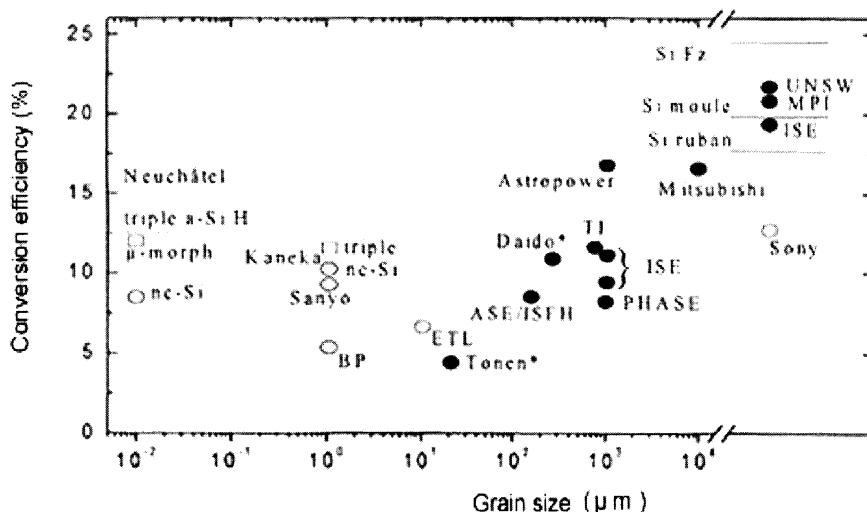


Fig. 5.6. Illustration of the laboratory cell efficiencies of thin-silicon-film technologies as a function of the grain size (from amorphous to epilayers) [29]

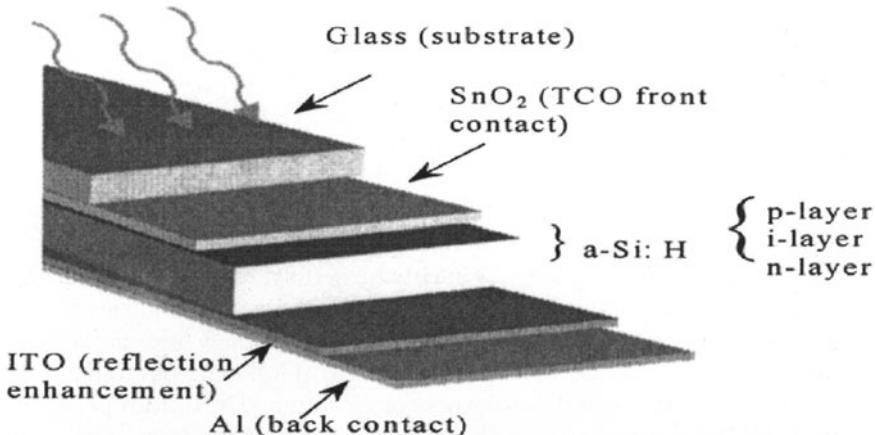


Fig. 5.7. Illustration of the amorphous-thin-silicon-film technology [30]

ious silicon film technologies, from amorphous to almost single-crystal films (quasi-monocrystalline epitaxial layers).

For amorphous silicon, which can be prepared at low temperatures, the substrate is usually glass. The collecting structure consists generally of three layers: a boron-doped and a phosphorus-doped layer with an intrinsic layer in between for improved charge transport in the internal electric field (p–i–n); see Fig. 5.7 [30].

In order to yield higher efficiencies, double and triple p–i–n structures have been developed with thinner layers, resulting in improved minority carrier diffusion length values. Also, small grain sizes in an otherwise amorphous matrix improve the charge transport properties, so that micro- or polymorphous materials have been developed successfully [31] (left part of Fig. 5.6).

The right side of Fig. 5.6 represents the results of research efforts either to growth by CVD processes epitaxial silicon on quasi-single crystals obtained from thermal crystallization of a double porous layer on a Si wafer [32] (see Fig. 5.8), or to transfer a silicon film from a silicon ingot by various techniques, including fragility enhancement of an intermediate layer and ion beam processing [33]. The data in the central region of Fig. 5.6, with grain sizes corresponding to polycrystalline silicon films, are not discussed in this chapter. See Chap. 4 on “Polycrystalline Silicon Films for Electronic Devices” for further information.

From an industrial point of view, the thin-film technologies are important both for reducing substantially the amount of silicon used and for being compatible with roll-to-roll processes [34] (Fig. 5.9) with flexible substrates such as stainless steel or plastic films.

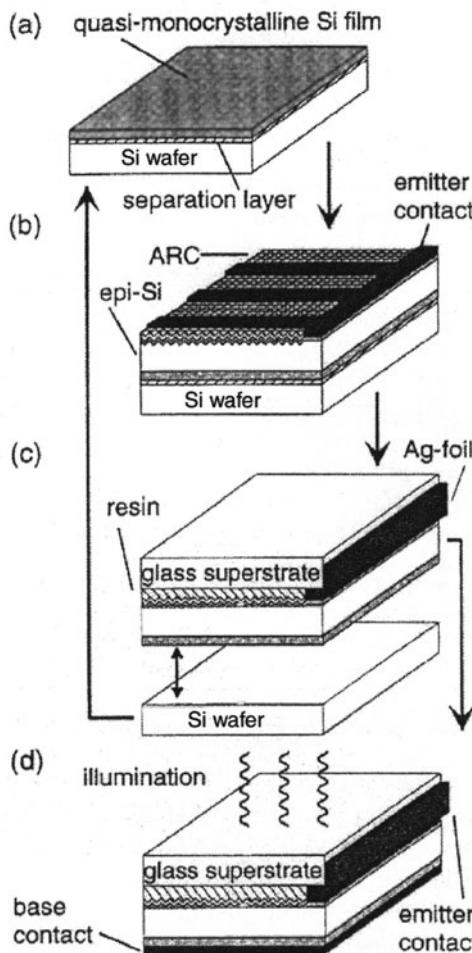


Fig. 5.8. Illustration of the transfer technology of quasi-monocrystalline silicon films [32]

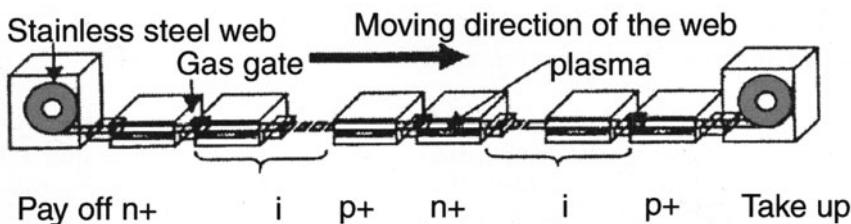


Fig. 5.9. Illustration of the roll-to-roll process adapted to thin-film technologies [34]

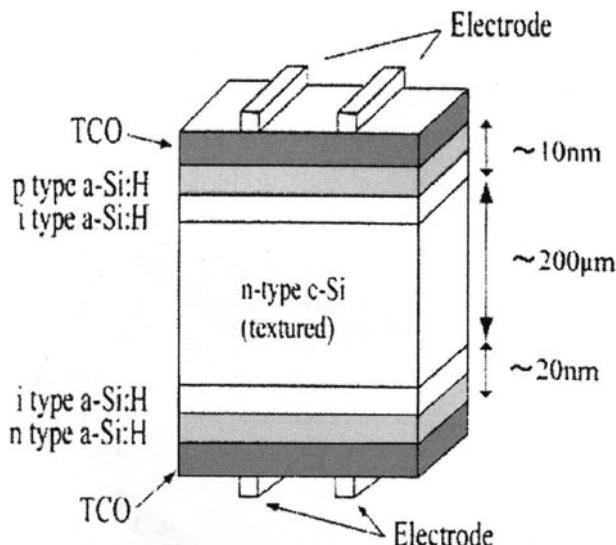


Fig. 5.10. Newly developed HIT structure, which combines the advantage of crystalline bulk transport properties with low-temperature amorphous heterostructures [35, 36]. TCO, transparent conductive oxide

(b) Wending of Amorphous with Crystalline Silicon for PV

Sanyo has developed a new structure (HIT) [35] (Fig. 5.10) which combines the efficient charge transport properties of bulk crystalline silicon with the low-temperature process of amorphous silicon deposition. Two heterostructures of amorphous silicon were produced on both sides of the cell. Efficiencies up to 21% have been obtained on n-type crystalline silicon with two-sided illumination [36]. If the industrial feasibility is clearly demonstrated for single crystals, the feasibility of the process on p-type multicrystalline substrates has to be confirmed next, before the process is used industrially on a larger scale.

5.3 Transport Properties in PV Silicon

The crystallographic properties of mc-Si materials are now quite satisfactory: a columnar structure with large grains more than 1 cm^2 in size, and few dislocations and intragrain defects. However, multicrystalline silicon contains larger quantities of impurities than does single-crystal silicon, which can have detrimental effects on the bulk minority carrier diffusion length ($L_{n,p}$). These impurities, which include metals as well as high concentrations of carbon and oxygen, can degrade the photovoltaic properties of solar cells and require specific treatments, presented in Sect. 5.3.2.

5.3.1 Effects of Defects and Impurities on the Transport Properties of Silicon

Impurities and residual defects can have detrimental effects on charge carrier transport and degrade the photovoltaic properties of a solar cell. Specific treatments such as gettering, applied separately or in conjunction with doping steps, can limit or avoid the degradation of the bulk diffusion length. The efficiency of these treatments depends strongly on the type of impurities present in the bulk.

Hydrogen passivation of electrically active defects is another efficient way to improve the transport properties, especially for materials having high densities of intragrain defects and smaller grains. It can be advantageously coupled with gettering processes, as well as with processes used to provide antireflection coatings such as SiN:H plasma deposition.

(a) Defects in Single-Crystal Silicon (sc-Si)

In high-efficiency cells fabricated using single-crystal silicon, a small degradation of the bulk minority carrier diffusion length $L_{n,p}$ can result from the presence of residual defects and impurities. It was found that for cells manufactured on boron-doped Cz silicon material, the efficiency decreases with time and with illumination. This effect, which does not appear in FZ silicon or in Ga-doped Cz materials, is now attributed to a vacancy–oxygen–boron complex [37, 38].

(b) Defects in Multicrystalline Silicon (mc-Si)

Multi-crystalline silicon wafers are characterized by a large grain size (larger than a few mm) and by a high density and large variety of intragrain defects, independently of the method used to grow the ingots, i.e. for different crucibles, liners, thermal steps during the solidification, etc. Dislocations, dislocation lineages, dislocation tangles, twins (decorated or not by dislocations) and subgrain boundaries are found. These defects can interact with grain boundaries [39].

Point defects such as vacancies and self-interstitials must also be taken into account, as well as impurities such as metal, oxygen and carbon atoms. These point defects and impurities can segregate at extended defects and form precipitates at defects sites or in the grains. The simultaneous presence of all these imperfections is the main characteristic of this type of materials, which appear to be very complex systems.

The interactions between defects and impurities are driven by chemical and mechanical potential gradients, since oxygen and carbon, at least, are frequently present in supersaturation concentrations [39]. The rate of interaction increases with temperature. Therefore the interaction is strongly influenced

by high-temperature processing steps during solar cell manufacture and explains the large variations in conversion efficiency; see the section devoted to solar cells (Sect. 5.4).

5.3.2 Improvement of the Material by Gettering

The main part of the recombination strength of extended defects is due to segregation of impurities. Impurity precipitates are also sources of recombination centres. Thus mc-Si wafers could be markedly improved by removing impurities from the bulk, especially metallic impurities.

The “gettering” effect can be used to clean the material. The basic mechanisms were divided into three categories by Schröter et al. [40]: (i) relaxation-induced gettering, i.e. trapping of supersaturated impurities by crystallographic defects; (ii) segregation-induced gettering by enhancement of the solubility in a certain region of the wafer; and (iii) injection-induced gettering, i.e. injection of self-interstitials and interaction with impurities.

(a) External Gettering

Multicrystalline silicon wafers are used exclusively to produce solar cells. Such wafers have the advantage of being cost-effective compared with single-crystal wafers. Hence inexpensive gettering techniques must be employed. The best solutions will be those which correspond to processing steps already included in the industrial cell fabrication process, such as phosphorus, boron or aluminium diffusion near the surfaces, and oxidation. “External” gettering of impurities is based on the creation of gettering sites at or near the external surfaces, and on the extraction, fast diffusion, and capture of impurities.

A heavy phosphorus diffusion near the surface is able to create crystallographic defects which behave as gettering sites. The phosphorus concentration is frequently beyond the solubility, and precipitates of SiP are formed. Ourmazd and Schröter [41] have shown that dislocations which trap fast diffusers such as Fe and Ni are generated at the border of these orthorhombic precipitates.

Gettering also results from the existence of a region where the impurity solubility is enhanced by the high phosphorus doping level, owing to the formation of complexes between impurities and doping atoms or to the shift of the Fermi energy level towards the conduction band. Highly N^+ phosphorous-doped regions, for instance, contain a high density of vacancies due to the phosphorus diffusion, which can trap interstitial atoms and increase their solubility.

Phosphorus diffusion has a further beneficial effect on gettering. External gettering requires the extraction of impurities from their substitutional positions, from precipitates, from segregation sites at extended defects or from the impurity cloud which surrounds these defects. The extracted impurities

must reach interstitial positions in order to diffuse rapidly through the crystal to the gettering sites. The extraction is facilitated when an excess of Si_i is injected into the bulk during phosphorus diffusion and participates in the well-known kick-out mechanism and in the shrinkage of precipitates [42]. This is exactly what phosphorus diffusion can do when SiP precipitates are formed near the surface, because this formation involves a molar-volume expansion which generates an excess of Si_i .

Very large minority carrier diffusion lengths could be obtained when thin wafers ($\approx 200 \mu\text{m}$ thick) of multicrystalline material were phosphorus diffused in a classical furnace at 900°C for several hours, using POCl_3 as a diffusion source [43]. The effect of phosphorus diffusion was applied successfully to various multicrystalline materials such as Silso, Polix, Eurosil and Solarex [44], increasing solar cell efficiencies [45–47] by the removal of metallic impurities such as Fe , Cu and Ni from the active part of the cell and by their accumulation in the front N^+ layer [43, 45].

A large number of reports show that the minority carrier diffusion length can also be enhanced if the silicon wafers are annealed in the presence of an aluminium layer. Aluminium is an acceptor in silicon, and its diffusion leads to P^+ -type regions, which can be used as ohmic contacts or to realize a back surface field (BSF). This diffusion length improvement, which can be as large as several hundred micrometers [48–50], suggests the existence of a gettering mechanism: metallic impurities could be trapped at the aluminium/silicon interface by an aluminium–silicon alloy. This process was found to be enhanced by dislocations.

Aluminium gettering is also an efficient means to improve multicrystalline silicon wafers. Large values of the effective diffusion length L_{eff} (frequently larger than $200 \mu\text{m}$) are obtained after the deposition of a $1 \mu\text{m}$ thick layer of aluminium and annealing at 950°C for 30 min, especially when the wafers have been previously gettered by phosphorus diffusion [51].

These improvements have been applied to solar cells to obtain high efficiencies [51, 52]. In spite of the complexity of the measurements (of the back surface field effect), a large part of the improvements of the cell efficiency have been attributed to gettering. Note that phosphorus and aluminium gettering can advantageously be applied simultaneously at the same temperature [49, 53].

(b) The Particular Case of Rapid Thermal Gettering

Despite its success in the microelectronics industry, rapid thermal processing (RTP) had to overcome a major difficulty for PV devices. RTP was revealed to be much more sensitive to contamination than were classical thermal treatments with slow cooling rates. In multicrystalline materials the higher concentrations of impurities are in supersaturation during the high-temperature step, and in the presence of a large density of crystallographic defects, some of

them can be frozen into electrically active sites during the quenching step [54] which is inherently associated with the fast cooling rate of an RTP cycle.

Fast-diffusing impurities (e.g. Co, Cu and Ni) generally have enough time to form electrically inactive precipitates, whereas the slower-diffusing impurities (e.g. Mo, Ti, V, Cr and Fe) mostly remain interstitially dissolved, depending on the quenching speed, and introduce recombination centres. For this reason the second group of metals is harmful to the performance of solar cells at much lower concentrations than the first group [55].

It has been shown that rapid thermal gettering can be just as efficient in improving the bulk minority carrier lifetime as conventional gettering [56]. The precise mechanism which allows gettering of metallic impurities within seconds is not yet well established. Some suggestions can be found in [57], where rapid thermal diffusion of phosphorus from a spin-on deposited source was applied to gold-contaminated monocrystalline silicon.

For mc-Si materials, simultaneous rapid thermal co-diffusion of phosphorus and aluminium [58] is required. Only under this condition does the cumulative effect of both gettering mechanisms lead to an improved minority diffusion length. The optimal process temperature depends on the mc-Si material, owing to the permanent competition between gettering and the quenching-induced formation of recombination centres; see also [59].

(c) Improvement by Passivation of the Material

Besides the remarkable efforts in research on impurity gettering as described above, many kinds of hydrogen passivation techniques have been developed in order to neutralize the activity of dangling bonds and of residual metallic impurities. It has been shown, in particular, that passivation by atomic hydrogen is very effective in improving the performance of multicrystalline-silicon solar cells [60,61].

Hydrogenation can be performed by various techniques, including RF, remote-plasma and ion beam processes. These treatments need to be performed on the finished cells since the bonds of hydrogen with defects or impurities can be broken at temperatures above 400°C [62], where hydrogen diffuses out of the wafer.

Hydrogen passivation can be applied to the front side and/or to the back. However, for all techniques involving energetic beams [63], back-surface irradiation is preferred in order to avoid damaging the active part of the cell. The most effective hydrogenation procedure was performed by Kyocera [64] during silicon nitride deposition (by PECVD) and is now widely used in the industry in order to form the antireflective coating simultaneously. By this method the cell is encapsulated and passivation can be performed during a short-time high-temperature sintering of the screen-printed contacts, with an additional positive effect on the bulk diffusion length if aluminium is present on the back surface [65].

(d) Role of Hydrogen in Amorphous Silicon

It has been proved that atomic hydrogen is very effective in improving the transport properties of amorphous materials by saturating the dangling bonds, and similarly in most materials with a low degree of crystallization. The hydrogenation is generally done during plasma or CVD deposition processes. The stability of this passivation process depends mostly on the deposition temperature. At 200°C, Si–H(n) bonds are formed, which are less stable than Si–H bonds formed close to 400°C [62] (for more details see Chap. 7, “Amorphous Hydrogenated Silicon”).

5.4 Silicon Solar Cells

5.4.1 Silicon Solar Cell Technology in Comparison with Other Technologies

Silicon solar cell technology has now reached almost its theoretical efficiency limit with the development of the PERL, passivated emitter rear locally-diffused cell by Green at the University of New South Wales (see Fig. 5.11) [66]. The technology using amorphous material which started in the 1960s has now been overtaken by the newly developed CIS, Cu-In-Se compounds. The Si thin-film technology now has to improve its crystal quality and efficiency. Organic cells, which have stability problems, still have a long way to go before industrial production. The new concepts which have been suggested to circumvent the thermodynamic efficiency limit of silicon cells are still lacking experimental confirmation (see Fig. 5.12) [67].

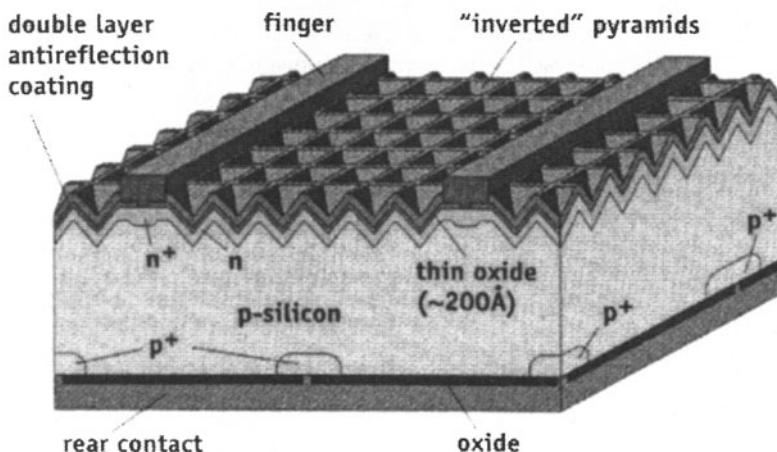


Fig. 5.11. Illustration of the best cell structure that can be formed on an FZ silicon substrate, confirming that silicon solar cell technology has now reached almost its efficiency limit with the development of the PERL cell by Green [66]

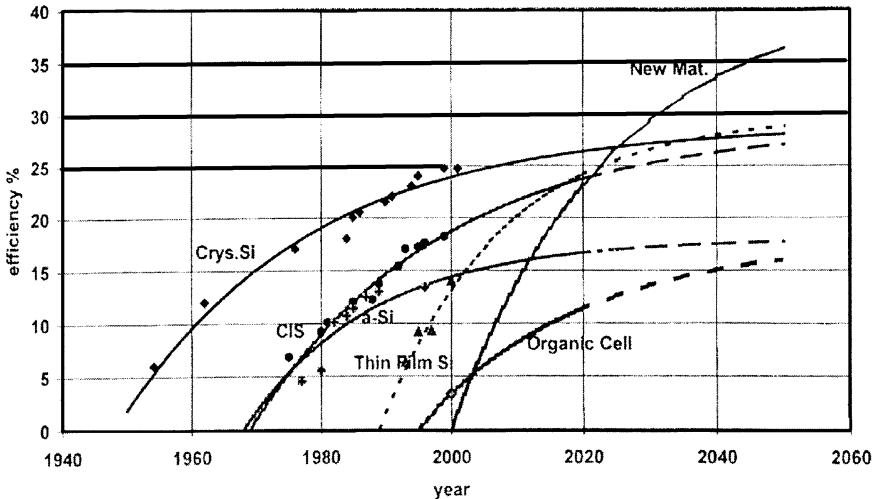


Fig. 5.12. Evolution of the laboratory conversion efficiencies for the three generations of silicon solar cells: the first generation, represented by crystalline silicon; the second one, ranging from amorphous silicon, CIS compounds, which have rapidly surpassed a-Si, and silicon thin films to the more recent organic cells; and the third one, with new materials and new concepts which may improve the efficiency over the present limit [67]

5.4.2 Multicrystalline Silicon Solar Cell Technology

Many organizations have been involved in research and development on multicrystalline silicon solar cells processed using one of the above-mentioned casting methods. The University of New South Wales [68] and Georgia Tech. [47,52] were the first to successfully achieve high conversion efficiencies in 1 to 4 cm^2 laboratory-scale multicrystalline-silicon solar cells.

The conversion efficiency of large, 100 cm^2 pre-manufacturable multicrystalline-silicon solar cells has rapidly reached a maximum of 17.2% by use of a mechanically grooved front electrode [69], screen-printing technology [70] and selective emitters [71]. Moreover, on larger-area (225 cm^2) solar cells, efficiencies up to 16.7% [72] have been achieved.

However, the industrial conversion efficiencies of 100 cm^2 cells in production are between 12.5 and 14.5% (see Fig. 5.13 [59]). For industrial cells in which $\text{SiN}(\text{H})$ has been introduced in production as an antireflective coating and as bulk and surface passivation, the efficiencies are closer to 16%.

For rapid thermal processing with fast cooling, it has been shown that the efficiencies of multicrystalline-silicon solar cells can be of the same order as those obtained by classical thermal processes [73]. The highest value, of 16.7% on a 25 cm^2 cell on Polix material from Photowatt, was obtained at the CNRS-PHASE Laboratory [74].

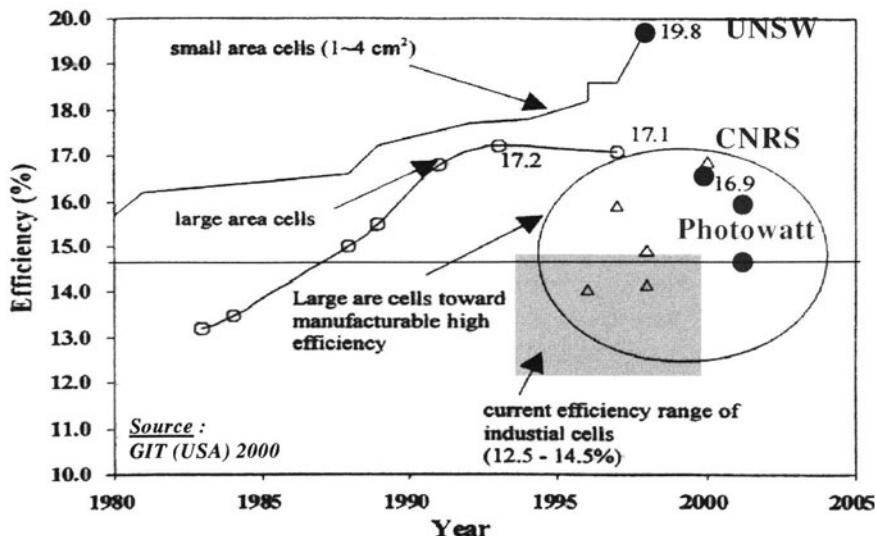


Fig. 5.13. Evolution of the laboratory conversion efficiencies for multicrystalline-silicon solar cells from small laboratory cells to the current industrial cells [59]

5.5 Conclusion

The strategy of future research in industrial silicon solar cell technology has to be concentrated upon the fabrication of larger cells (more than the present $15 \times 15 \text{ cm}^2$), cut from larger cast multicrystalline ingots or from continuous-pulling electromagnetic-casting materials. Oxygen, carbon and metallic impurity concentrations need to be reduced and, consequently, bulk transport properties enhanced. Classical and rapid thermal gettering as well as hydrogenation processes can still be improved, reducing the difference in the conversion efficiency between multi- and monocrystalline silicon cells.

Solar cells have also to be made thinner ($150 \mu\text{m}$ slices for the near future and down to $50 \mu\text{m}$ if possible) in order to reduce the silicon losses during ingot slicing from 40 to 20% and to remain competitive with the ribbon and thin-film technologies for another 50 years.

References

1. *Eurec Position Paper*, Eurec Agency, Brussels, BE-1040 (2000)
2. *World Energy Outlook 2003*, IEA Annual Report
3. N. Yuge, M. Abe, K. Hanazawa, H. Babe, N. Nakamura, Y. Kato, Y. Sakaguchi, S. Hiwasa, M. Obashi: 11th Int. Photovoltaic Science & Engineering Conf. (Sept. 20–24, 1999), Sapporo, Japan, p. 115
4. B.H. Authier: German Patent (DOS) No. 25 0883 (1975); H. Fischer: Proc. 1st EC PV SEC, Luxembourg (1977) p. 52

5. G.M. Storti: Proc. 15th IEEE PVSC, Kissimmee (1981) p. 442
6. E. Schmid, C.P. Khattak: Proc. 5th EC PV SEC, Athens (1983) p. 1019
7. J. Fally, D. Guignot, L. Goeffron: Proc. 7th EC PV SEC, Seville (1986) p. 754
8. T.E. Ciszek, G.H. Schwuttke, K.H. Young: *J. Cryst. Growth* **46**, 527 (1979)
9. T. Saito, A. Shimura, S. Ichikawa: Proc. 15th IEEE PVSC, Kissimmee (1981) p. 576
10. D. Margadonna, F. Ferrazza, R. Peruzzi: Proc. 10th EC PV SEC, Lisbon (1991) p. 678
11. I.A. Dorrity, B.J. Garrard, D.A. Hukin: Proc. 10th EC PV SEC, Lisbon (1991) p. 317
12. W. Koch, W. Krumbe, H. Lange, W. Schmidt, F. Schomann, G. Wahl: Proc. 12th EC PV SEC, Amsterdam (1994) p. 797
13. Photowatt International S.A., FR-38300 Bourgoin-Jallieu
14. K. Kaneko, T. Misawa, K. Tabata: Proc. 21st IEEE PVSC, Kissimmee (1990) p. 674
15. T.F. Ciszek: *J. Electrochem. Soc.* **132**, 963 (1985)
16. K. Kaneko, R. Kawamura, T. Misawa: Proc. 1st WCPEC, Hawaii (1994) p. 30
17. I. Perichaud, G. Dour, F. Durand, D. Sarti, G. Goaer, Q.N. Le, F. Floret, S. Martinuzzi: Proc. 13th EC PV SEC, Nice (1995)
18. EMIX S.A., FR-23300 La Souterraine
19. D.A. Hukin: Proc. 4th PVSEC, Sydney (1989) p. 719
20. K. Kaneko, K. Sasatani, M. Ohnishi, N. Kimura: Proc. 11th Int. PVSEC Conf., Sapporo, Japan (1999) p. 119
21. R.G. Seidensticker, L. Scudder, H.W. Brandhorst: Proc. 11th IEEE PVSC, Scottsdale, AZ (1975) p. 299
22. ASE-Americas, Inc., Billerica, MA, 01821-3980, USA
23. K.V. Ravi, H.B. Serreze, H.E. Bates, A.D. Morrison, D.N. Jewett, J.C.T. Ho: Proc. 11th IEEE PVSC, Scottsdale, AZ (1975) p. 281
24. A. Lesk, A. Baghdadi, R.W. Gurtier, R.J. Ellis, J.A. Wise, M.G. Coleman: Proc. 12th IEEE PVSC, Baton Rouge (1976) p. 173
25. C. Belouet, J.J. Brissot, R. Martres, N.T. Phuoc: Proc. 1st EC PV SEC, Luxembourg (1977) p. 164
26. EurObserv'ER, Systèmes Solaires, No. 136 (2000)
27. B.H. Mackintosh, M.P. Ouellette, M.D. Rosenblum, J.P. Kaleijs, B.P. Piwczyk: 28th IEEE PVSC, Anchorage, AK (2000) p. 46
28. RWE-Schott-Solar, Alzenau, Germany
29. R.B. Bergmann: *Appl. Phys. A* **69**, 187 (1999)
30. P. Roca i Cabarrocas, Ecole Polytechnique, Palaiseau, France, personal communication
31. P. Roca i Cabarrocas, A. Fontcuberta i Morral, Y. Poissant: *Thin Solid Film* **403–404**, 39 (2002)
32. R.B. Bergmann, T.J. Rinke, C. Berge, J. Schmidt, J.H. Werner: 12th Int. PVSEC Conf., Jeju, Korea (2001) published in: *Solar Energy Materials and Solar Cells*, **74**, 213 (2002)
33. P.J. Riberon, A. Beaumont, A. Laugier, A. Kaminski, A. Fave, M. Lemiti, G. Fantosi: ADEME-CNRS Annual Seminar on PV Materials and Processes, Sophia-Antipolis (2002) p. 111
34. M. Sano, K. Saito, S. Okabe, S. Sugiyama, A.K. Ogawa: Proc. 12th Int. PVSEC Conf., Jeju, Korea (2001) p. 29

35. T. Sawada, N. Terada, S. Tsuge, T. Baba, T. Takahama, K. Wakisaka, S. Tsuda, S. Nakano: 1st WCPEC, Hawaii (1994) p. 1219
36. M. Taguchi et al: Prog. Photovolt. Res. Appl. **8**, 503 (2000)
37. S.W. Glunz, S. Rein, W. Warta, J. Knobloch, W. Wettling: Proc. 11th Int. PVSEC Conf., Sapporo, Japan (1999) p. 549
38. J.Y. Lee, S. Peters, S. Rein, S.W. Glunz: Proc. 12th Int. PVSEC Conf., Jeju, Korea (2001) p. 27
39. J.C. Muller, S. Martinuzzi: J. Mater. Res. **13**, 2721 (1998)
40. W. Schröter, M. Seibt, D. Gilles: Mater. Sci. Technol. **4**, 540 (1992)
41. A. Ourmazd, W. Schröter: Appl. Phys. Lett. **45**, 781 (1984)
42. J.S. Kang, D.K. Schroder: J. Appl. Phys. **65**, 2974 (1989)
43. I. Péricaud, S. Martinuzzi: J. de Phys. III **2**, 313 (1992)
44. I. Péricaud, F. Floret, M. Stemmer, S. Martinuzzi: Solid State Phenomena **32**, 77 (1993)
45. B. Sopori, L. Jastrzebski, T. Tan, S. Narayanan: Proc. 12th EC PV SEC, Amsterdam (1994) p. 1003
46. L.Q. Nam, M. Rodot, M. Ghannam, D. Sarti, I. Péricaud, S. Martinuzzi: Int. J. Solar Energy **11**, 273 (1992)
47. A. Rohatgi: Proc. 23rd IEEE PV SEC, Louisville (1993) p. 52
48. R. Sundaresan, D.E. Burk, J.G. Fossum: J. Appl. Phys. **55**, 1162 (1984)
49. L.A. Verhoef, S. Roorda: Proc. 20th IEEE PVSC, Las Vegas (1988) p. 1551
50. O. Porre, M. Pasquinelli, S. Martinuzzi, I. Péricaud: Proc. 11th EC PV SEC, Montreux (1992) p. 1053
51. M. Pasquinelli, S. Martinuzzi, J.Y. Natoli, F. Floret: Proc. 22nd IEEE PVSC, Las Vegas (1991) p. 1035
52. A. Rohatgi, P. Sana, J. Salami: Proc. 11th EC PV SEC, Montreux (1992) p. 159
53. L.A. Verhoef, P.P. Michiels, S. Roorda, R. Sinke, R.J. Van Zolingen: Mater. Sci. Eng. B **7**, 49 (1990)
54. W. Eichhammer, Vu-Thuong-Quat, P. Siffert: J. Appl. Phys. **66**, 3857 (1989)
55. E. Weber. In: *Impurity Diffusion and Gettering in Silicon*, Mater. Res. Soc. Proc. **36**, 3 (1985)
56. B. Hartiti, A. Slaoui, J.C. Muller, R. Stuck, P. Siffert: J. Appl. Phys. **71**, 5474 (1992)
57. B. Hartiti, A. Slaoui, J.C. Muller, P. Siffert: Appl. Phys. Lett. **63**, 1249 (1993)
58. B. Hartiti, A. Slaoui, J.C. Muller, P. Siffert, R. Schindler, I. Reis, B. Wagner, A. Eyer: Proc. 23rd IEEE PV SEC, Louisville, (1993) p. 224
59. A. Rohatgi, V. Yelundur, J. Jeong, A. Ristow, A. Ebong: 10th Workshop on Crystalline Silicon Solar Cell Materials and Process, (Copper Mountain, CO August 2000), p. 12
60. J.C. Muller, E. Hussian, P. Siffert, D. Sarti: Proc. 9th EC PV SEC, Freiburg (1989) p. 407
61. B. Sopori, K. Jones, X. Dung, R. Matson, M. Al-Jassim, S. Tsuo, A. Doolittle, A. Rohatgi: Proc. 22nd IEEE PVSC, Las Vegas (1991) p. 833
62. J.C. Muller, B. Hartiti, E. Hussian, J.P. Schunck, P. Siffert, D. Sarti: Proc. 22nd IEEE PVSC, Las Vegas (1991) p. 883
63. S. Sivoththaman, M. Rodot, J.C. Muller, B. Hartiti, M. Ghannam, H.E. El-gamel, J. Nijs, D. Sarti: Appl. Phys. Lett. **62**, 3172 (1993)
64. H. Watanabe: 4th Int. PVSEC Conf., Sydney (1989) p. 103

65. V. Yelundur, A. Rohatgi, J.W. Jeong, A.M. Gabor, J.I Hanoka, R.L. Wallace: Proc. 28th IEEE PVSC, Anchorage, AK (2000) p. 91
66. M.A. Green, A. Wang, G.F. Zheng, Z. Zhang, S.R. Wenham, J. Zhao, Z. Shi, C.B. Honsberg: Proc. 12th EC PVSEC, Amsterdam (1994) p. 776
67. A. Goetzberger, J. Luther, G. Willeke: Proc. 12th Int. PVSEC Conf., Jeju, Korea (2001) p. 5
68. S. Narayanan, S.R. Wenham, M.A. Green: Proc. 4th PV SEC, Sidney (1989) p. 111
69. H. Nakoya, M. Nishida, Y. Takeda, S. Moriuchi, T. Tonegawa, T. Machida, T. Nunoi: 7th PV SEC, Nagoya (1993) p. 91
70. J. Copye, J. Szlufcik, H. Elgamel, M. Ghannam, P. De Schepper, J. Nijs, R. Mertens: Proc. 22nd IEEE PVSC, Las Vegas (1991) p. 873
71. J. Szlucik, F. Duerinckx, J. Horzel, E. Van Kerschaver, S de Wolf, P. Choulat, H. Dekkers, J. Nijs: Proc. 12th Int. PVSEC Conf., Jeju, Korea (2001) p. 305
72. K. Shirasawa, H. Takahashi, Y. Inomata, K. Fukui, K. Okache, M. Takayama, H. Watanabe: Proc. 12th EC PV SEC, Amsterdam (1994) p. 757
73. S. Sivoththaman, B. Hartiti, J. Nijs, A. Barhdadi, M. Rodot, J.C. Muller, W. Laurey, D. Sarti, Proc. 12th EC PV SEC Amsterdam (1994) p. 47
74. S. Noël, H. Lautenschlager, J.C. Muller: Prog. Photovolt. Res. Appl. **9**, 41 (2001)

Part III

Epitaxy, Films, and Porous Layers

6 Films by Molecular-Beam Epitaxy

I. Eisele, J. Schulze, E. Kasper

6.1 Equipment Principles and Growth Mechanisms

Epitaxial growth proceeds by the attachment of atoms on the correct positions of a given lattice. The orientation and atomic spacings of the lattice are usually those predicted from a single-crystalline substrate with a clean surface. The attachment of atoms from the vapour phase typically follows a three-step scheme of adsorption, diffusion and incorporation into surface steps (Fig. 6.1). The adsorbed atoms, called adatoms, are in a precursor state for later incorporation into the lattice. The adsorption energy W_{ad} is lower than the binding energy W_b , usually 1/2 to 2/3 of W_b .

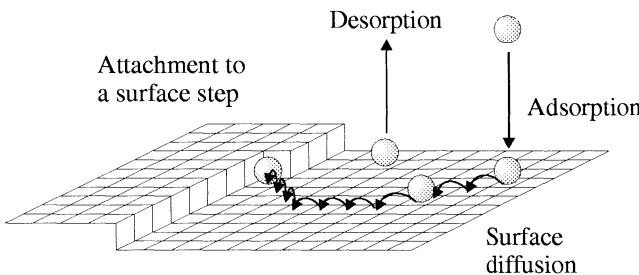


Fig. 6.1. Basic steps in the growth of an epitaxial layer from the vapour phase

But even when adsorbed, the adatom may escape by a later desorption step caused by thermal vibrations. A regular network of places on the surface is available for the adatoms. The adatoms jump rather easily from one of these places to another one; this is described as surface diffusion with an activation barrier. By a random walk, if they are not desorbed in the meantime the adatoms will reach a step nearby and become incorporated into the crystal (Fig. 6.2).

The most common source of monatomic steps, with a step height h , in silicon is the unintentional misorientation i of commercial wafers, with the angles i being typically between 0.1° and 0.5° . Even with nominally oriented substrates, terrace widths L of 15 nm to 75 nm are expected.

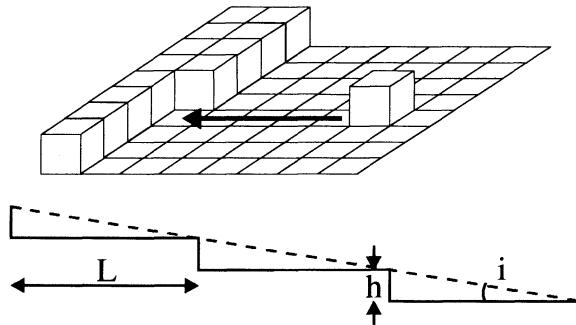


Fig. 6.2. Step growth by the attachment of diffusing adatoms. The distance between steps (the terrace width L) is given by the misorientation i of the substrate surface and the step height h

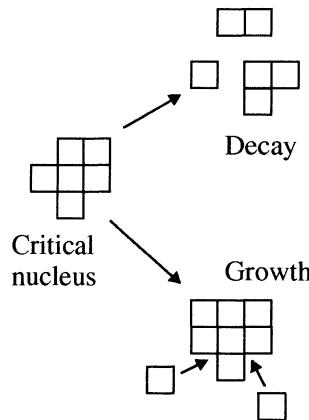


Fig. 6.3. Nucleation of adatoms. The size of the critical nucleus is dependent on supersaturation because of the dynamic equilibrium between decay and growth processes

When all adatoms reach the already existing misorientation steps (in imperfect substrates, dislocation steps also act in a similar manner) then the monatomic steps move laterally forward by adatom capture. This happens at higher temperatures; as a rough estimate one can consider temperatures above $T_m/2$ (T_m is the melting point, 1734 K for Si). At lower temperatures the slowly moving adatoms nucleate into two-dimensional islands (Fig. 6.3).

A critical nucleus is defined by a size where growth by capture of adatoms is more probable than decay of the nucleus. With higher supersaturation the size of the critical nucleus becomes smaller, being two atoms under most MBE conditions. Fig. 6.4 compares both adatom capture mechanisms in the

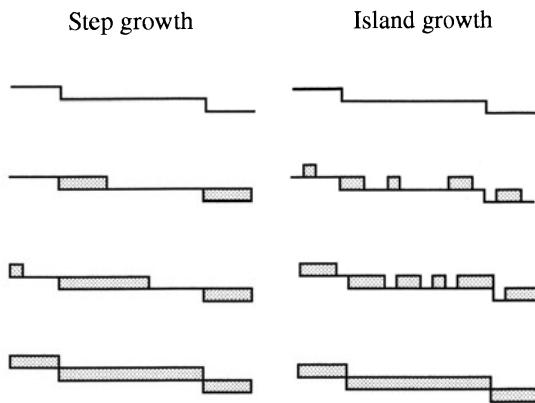


Fig. 6.4. Side view of the basic two-dimensional growth mechanisms

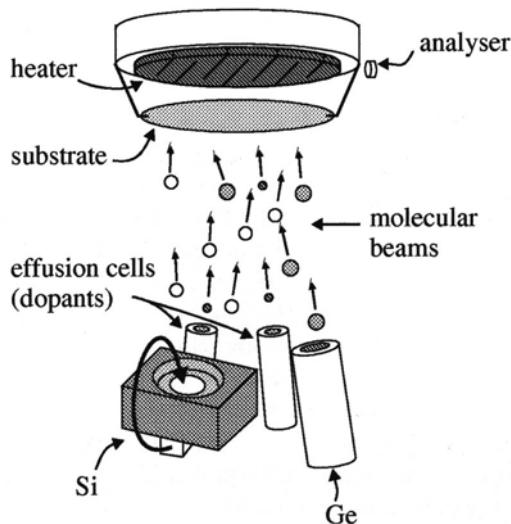


Fig. 6.5. Basic scheme of MBE

two-dimensional (van der Merwe) growth mode. The step flow mechanism is based on the lateral movement of pre-existing misorientation steps; the two-dimensional (2D) nucleation mechanism creates steps by a nucleation process. These nuclei annihilate after one monolayer so that 2D nucleation is a periodic process.

The molecular-beam epitaxy (MBE) process is performed in a very clean ultrahigh vacuum (UHV) environment. Atomic or molecular beams of the necessary species are directed towards the heated substrate and grow into an epitaxial layer as depicted in the principal Figs. 6.1 to 6.5 before. The

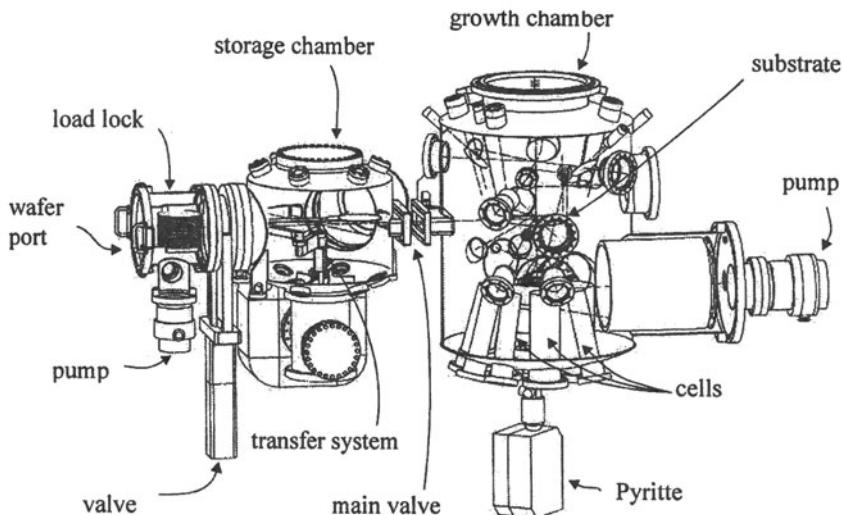


Fig. 6.6. Technical implementation of a Si MBE apparatus

molecular beams are created from heated source materials in effusion cells (Fig. 6.5). For highly refractory or reactive materials such as silicon, electron evaporators are used. Gas source MBE (GS-MBE) uses gaseous compounds (mainly compounds with hydrogen, or chlorine or organometallic compounds) directed by nozzles towards the substrate. Typical source materials contain the matrix elements Si, germanium (Ge) and carbon (C), and the dopant elements boron (B), gallium (Ga), antimony (Sb) and phosphorus (P); but also metals (silicides) and insulators (CaF_2 and oxides) can be grown by MBE. The MBE system is always equipped with one or more pieces of sophisticated in situ monitoring equipment.

In order to improve vacuum quality and to allow cluster processing, the growth chamber (Fig. 6.6) is connected to a storage chamber and a load lock, with additional options to cluster the chamber with analytical and processing equipment by a transfer system.

6.2 Historical Development

More than 35 years ago Unvala [1] and Hale [2] prepared and evaluated silicon epitaxial films grown in vacuum, in which they used some of the techniques now common in Si MBE. However, because of the insufficiently high vacuum conditions available at that time for such processes, this work may not strictly be considered to be the origin of Si MBE. In 1968, silicon was grown under typical conditions applicable to Si MBE, and Abbink et al. [3] grew epitaxial films with very smooth surfaces at temperatures of 800°C.

The significant details of Abbink et al.'s procedure include electron gun evaporation of the source material, beam modulation by a shutter, *in situ* monitoring of beam intensity and gas composition, and *in situ* cleaning of the substrate surface. Using a combination of *in situ* replication and transmission electron microscopy, they established an intrinsic 2D growth mode of clean silicon surfaces with fast-moving adatoms. These early investigations brought about an exciting amount of fundamental understanding of growth process on an atomic level [4–12].

Despite the excellent nature of the early work, it failed to have the necessary impact on silicon technology because of the lack of effective doping methods in vacuum deposition. The difficult problem of incorporating foreign atoms in a perfect silicon matrix is now understood, but for the early investigators a lack of understanding of the doping incorporation prevented useful practical application.

Heavy surface segregation of dopant atoms has been investigated since about 1975 and has been overcome by a variety of doping techniques. The first successful tests of MBE layers in integrated circuit (IC) fabrication [13] followed in 1985. For an overview and a bibliography of the relevant contributions up to 1985, see the first book on Si MBE [14].

MBE stimulated strongly the research on silicon-based heterostructures, both with insulators (e.g. CaF_2) and metals (e.g. silicides) and with other semiconductors (e.g. SiGe/Si). The latter heterostructure is now in common use for SiGe heterobipolar transistors (HBTs) in high-frequency circuits and will soon have broad utilization in “strained silicon MOSFETs” with sub-90 nm gate lengths.

Early attempts at vacuum evaporation of Si/Ge started at the end of the 1960s and culminated in 1975–1979 with a basic understanding of critical thicknesses and the formation of SiGe superlattices by Kasper [15, 16].

The reduction of growth temperatures to 550°C by Bean et al. [17] allowed higher critical thicknesses and started the second wave of strained-layer epitaxy on silicon.

6.3 Stability of Strained Heterostructures

The lattice constant of SiGe is slightly larger than that of Si. For a rough overview, the lattice mismatch f of an alloy can be linearly interpolated (Vegard's law) between the two parent materials:

$$f = \frac{a_f - a_s}{a_s} = 0.042 X, \quad (6.1)$$

where a_f and a_s are the lattice constants of the film and substrate, respectively, and X is the Ge content of the alloy.

For more exact calculations, a small parabolic deviation has to be considered. Recent investigations [18] on epitaxial material confirmed the old

measurements on bulk crystals [19]. The structural and morphological stability of lattice-mismatched heterostructures is rather complex and covers both extremes, instability and perfect stability. For device processing, a basic insight into the fundamental mechanisms of mismatch accommodation is necessary. Nature gives several answers to lattice mismatch. In the case of SiGe on Si they are a strained film, strain relaxation by misfit dislocations, surface undulations and cracking. Cracking happens only during cooling down of thick Ge films and need not be considered for most devices. To simplify the picture we first treat the equilibrium answer and then discuss the kinetic limitations imposed by the low growth temperatures and by processing of multilayer device structures.

6.3.1 Critical Thickness of Strained Layers

Up to a critical thickness t_c the layer is strained; this means that every atomic row in the substrate is continued across the interface. This low-thickness regime is completely stable. The reason is given by the energy balance, favouring strain in a small volume instead of creating possible atomic defect structures such as misfit dislocations. Misfit dislocations are line defects where atomic rows are not continued across the interface. A misfit dislocation is characterized by its line vector l and its Burgers vector b , which defines the inhomogeneous strain field around the dislocation. The glide plane given by the vectors l and b allows easy movement of the dislocation. In diamond lattices the glide plane is the densely packed (111) plane. Dislocations cannot end in the material; they either are closed or end at surfaces.

The equilibrium critical thickness may be calculated rather easily from basic dislocation theory when – the reader should keep this in mind – the film is assumed to be flat. The differences in published numerical values are caused by differing assumptions about the dislocation core and the range of the inhomogeneous strain [20]. The calculated equilibrium thickness t_{cm} (the index “m” refers to van der Merwe and to Matthews and Blakeslee, who pioneered this kind of equilibrium calculation) is rather small, e.g. 6 nm for a 25% SiGe alloy ($f \cong 0.01$):

$$\left(\frac{t_{cm}}{b} \right) f = 5.78 \times 10^{-2} \ln \left(\frac{t_{cm}}{b} \right). \quad (6.2)$$

The magnitude b of the Burgers vector is 0.384 nm in Si.

The concept of the equilibrium critical thickness in the simple form given is only meaningful for low lattice mismatches. At higher mismatch values the film strain is reduced by surface corrugations, which cause a conversion from the flat van der Merwe growth mode to direct island growth (Volmer–Weber mode) or island growth on a thin wetting layer (Stranski–Krastanov mode); for example, Ge on Si builds a wetting layer only about 0.5 nm thick, on which nucleation of islands takes place. When the islands grow larger, then misfit

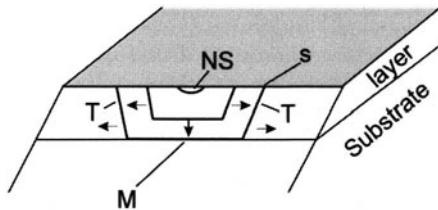


Fig. 6.7. Nucleation of dislocation half-loops (NS) and movement to the interface, creating segments M and T

dislocations are generated additionally. The generation of at least the first dislocations is assumed to be from surface nucleation sites (NS), which are either small imperfections or surface steps (Fig. 6.7). Under the strain force, dislocation half-loops move to the interface, creating a straight misfit dislocation segment (M) and two threading dislocation arms (T) connecting the misfit dislocation with the surface. A misfit dislocation network, when placed below an active device, could be the ultimate near-device getter, but the threading dislocations cause electronic problems with this kind of material.

6.3.2 Metastable Pseudomorphic Growth

The growth temperature of SiGe is strongly reduced (typically 450°C–750°C) compared with Si epitaxy. Surface atom migration and dislocation nucleation are kinetically suppressed. Flat, strained layers (pseudomorphic structure) are obtained under metastable growth conditions to much higher thicknesses and higher mismatch values than at equilibrium. In Fig. 6.8, the critical thicknesses at equilibrium (lower line) and those obtained at low growth temperatures (MBE, 550°C) are compared.

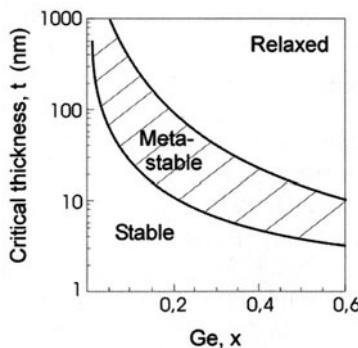


Fig. 6.8. Critical thickness t_c as function of the Ge content x . The stable region (6.2) and metastable region (6.3) are strained. The relaxed region contains misfit dislocations

The metastable critical thickness t_{cb} (the index “b” refers to People and Bean, who discovered the large amount of metastability and published a fit relation in 1984) is roughly given by

$$\left(\frac{t_{cb}}{b}\right) f^2 = \frac{1}{200} \ln \left(\frac{t_{cb}}{b}\right). \quad (6.3)$$

With growth temperatures below 550°C, even higher critical thicknesses can be obtained (ultrametastability).

6.3.3 Processing and Annealing of Device Structures

Strained layers with thicknesses below the equilibrium critical thickness t_{cm} are inherently stable. But metastable layers also turned out to be stable against heat treatments considerably higher than their growth temperature. The main reason is the silicon cap on top of the SiGe structure, which is grown either for functional purposes (HBT emitter) or to facilitate processing steps (oxidation and resist coverage). The equilibrium critical thickness is doubled by the cap, but the kinetic limitation is even stronger, as one can easily understand that, starting from a surface nucleation site, the dislocation half-loops are not forced to move through the unstrained cap. Processing of strained SiGe up to 850°C has been successfully demonstrated.

For processing with a low thermal budget, transient enhanced diffusion (TED) is gaining in importance, and this is nothing specific to heterostructures. But in heterostructures, the role of carbon doping in the suppression of boron diffusion [21] has been emphasized. Interstitials react with substitutional carbon and thereby avoid enhanced interstitialcy-driven boron diffusion – sharp, extremely highly doped regions are possible [22].

6.4 Dopant Distribution in Films Grown by Silicon MBE

6.4.1 The Doping Problem

Conceptually, and in practice, the easiest way to dope films during physical vapour deposition is co-evaporation, i.e. the dopant atoms are deposited together with the matrix atoms. In principle this allows almost arbitrary dopant concentration profiles. However, this is not true for Si MBE, because at typical growth temperatures $> 400^\circ\text{C}$ dopant incorporation is governed by severe surface segregation and low incorporation probabilities at the growing surface (Fig. 6.9). This well-known phenomenon leads to significantly broadened doping profiles and causes a severe “Si doping problem” [23–25].

The tendency of a dopant to segregate is characterized by a segregation coefficient r , which basically denotes the ratio between the deposited and the actual areal density of one monolayer [26].

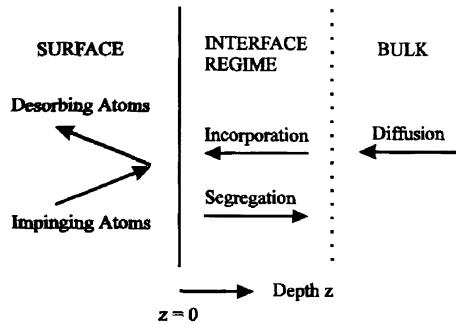


Fig. 6.9. Schematic diagram of the growing surface during Si MBE

As can be seen from Fig. 6.10, segregation can be reduced significantly by lowering the growth temperature. In addition, a decreasing flux of the impinging Si atoms lowers the transition temperature between thermodynamic equilibrium and the kinetically limited segregation regime [27].

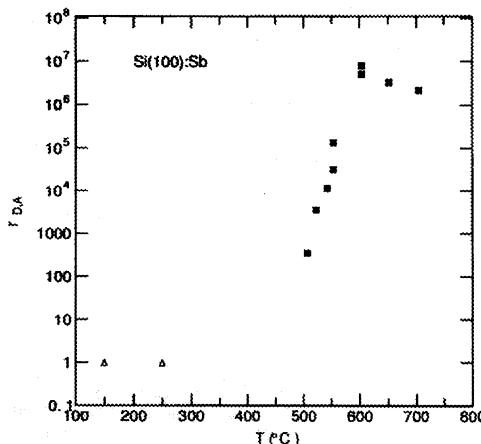


Fig. 6.10. Segregation coefficient r as a function of growth temperature for Sb doping. The *squares* correspond to experimental results from [27] for a growth rate of 0.3 nm/s ; the *triangles* are experimental data from [28] for 0.02 nm/s

However, arbitrarily lowering the growth temperature is not possible because there exists a “critical temperature” which separates the epitaxial, single-crystalline regime from the amorphous growth regime [29].

This model has been extended by Eaglesham [30]. For decreasing growth temperatures, he found a decreasing critical thickness t_{epi} for which single-crystalline growth is still possible. Above this thickness value, amorphous growth takes place. t_{epi} decreases exponentially with temperature (see

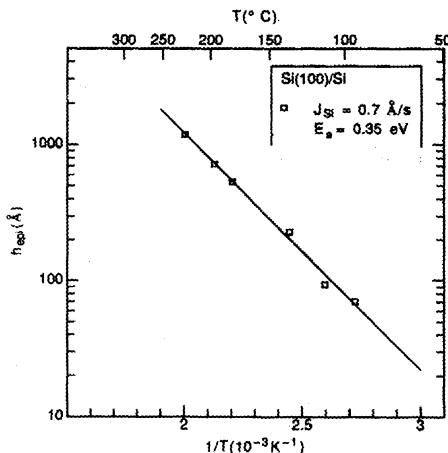


Fig. 6.11. Critical epi thickness for (100) Si at a growth rate of 0.07 nm/s as a function of growth temperature. The activation energy is indicated [26]

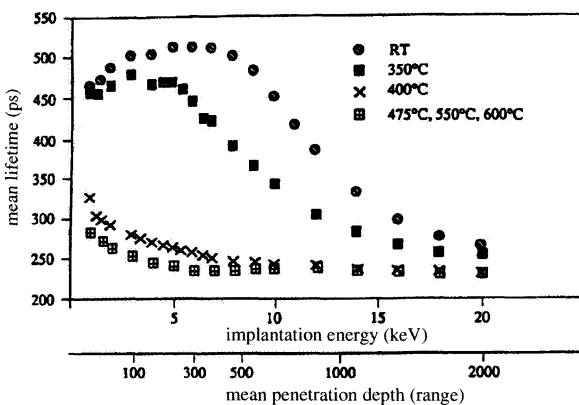


Fig. 6.12. Mean positron lifetime of MBE layers grown at various temperatures with a growth rate of 0.1 nm/s

Fig. 6.11). Increasing the growth rate also decreases the critical epi thickness. The transition between crystalline and amorphous growth is explained by an increasing surface roughness due to local nucleation.

For device applications, additional investigations regarding the defect density of single-crystal layers are of importance. With decreasing temperature, the surface diffusion of impinging atoms also decreases and eventually not all lattice sites are occupied, resulting in an increased concentration of point defects. This effect has to be avoided because these point defects can enhance diffusion effects and thus broaden the doping profiles. In addition, they act as recombination centres degrading the electrical device characteristics. Us-

ing positron annihilation spectroscopy, the resulting point defect density can be determined from the mean positron lifetime [31,32]. An increased lifetime can be correlated with an increased point defect density. A typical example for 800 nm thick MBE layers grown at various temperatures is shown in Fig. 6.12.

Below 475°C the enhanced lifetime is caused by an increased point defect density, whereas above this temperature the values correspond to the bulk values of a Si substrate and prove perfect crystalline quality. It also has been shown that the temperature for perfect growth can be lowered if intermediate annealing steps at temperatures above 500°C are carried out. For the fabrication of abrupt doping profiles and δ -type profiles, therefore, either low-temperature epitaxial growth or amorphous growth followed by annealing steps for recrystallization (solid phase epitaxy) can be used.

6.4.2 Abrupt and δ -Type Doping Profiles

δ -type doping profiles in silicon were fabricated for the first time by the solid phase epitaxy approach in 1987 by Zeindl et al. [33]. An areal dopant density of Sb was deposited on a Si surface at room temperature and, after growing a 3 nm thick amorphous cap layer of Si, recrystallization at 700°C was carried out. By this procedure, doping concentrations up to $2 \times 10^{14} \text{ cm}^{-2}$ were confined within about 1–2 nm. δ -layers have been investigated by numerous analytical techniques [26,34]. As an example, a high-resolution TEM (transmission electron microscopy) picture of Sb deltas is presented in Fig. 6.13.

Besides Sb, δ -layers of P [25], Ga [35] and B [36] were reported in the following years and numerous publications followed. For a review, see [26,34].

Low-temperature MBE of silicon and solid phase epitaxy turned out also to be the key to the fabrication of abrupt and nearly arbitrary doping profiles. In addition, secondary silicon ions, which are always present if silicon is evaporated by an e-gun, can be used in order to achieve well-defined doping profiles. If these ions are accelerated towards the substrate, they cause knock-on effects and the doping atoms are pushed into the bulk, thus reducing segregation effects [37]. Using one of these techniques or a combination of them, unsurpassed square profiles with a steepness of only a few nm per decade can be achieved by MBE.

Regarding the electronic properties of a δ -layer, it is of particular interest that the electronic levels form a set of two-dimensional subbands in a V-shaped potential. By solving self-consistently the one-dimensional Schrödinger equation for motion perpendicular to the surface and Poisson's equation in order to obtain the electrostatic potential, the energy diagram in Fig. 6.14 can be obtained. The energy separation between the subbands depends on the δ -doping concentration and can be as large as 150 meV. This means that inter-subband transitions and resonant tunneling effects can be considered for device applications.

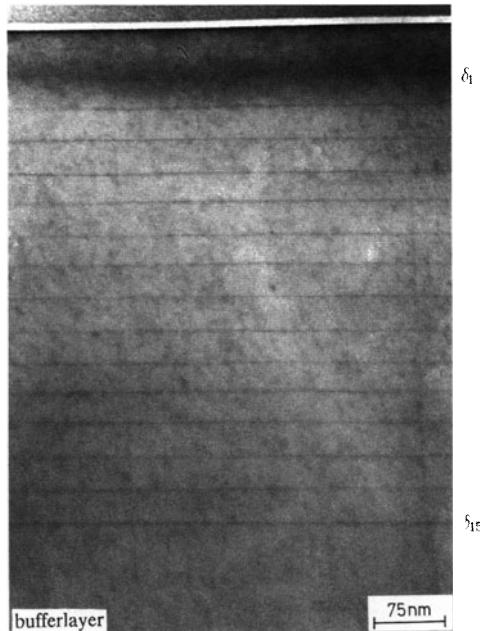


Fig. 6.13. Transmission electron micrograph of 15 Sb δ -layers spaced by 50 nm [38]

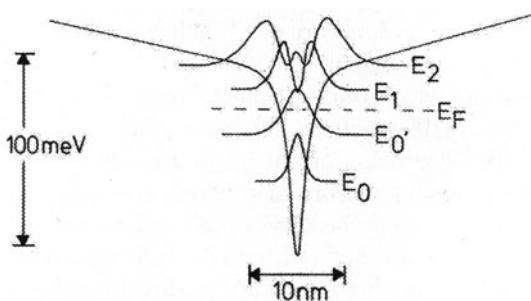


Fig. 6.14. Quantized energy levels, Fermi level E_F and electron charge distribution of the subbands for a δ -layer with $2 \times 10^{13} \text{ cm}^{-2}$ Sb atoms [39]

6.5 Semiconductor Device Research

MOS field effect transistors (MOSFETs) along with bipolar transistors, are the most popular devices in the commercial electronics market. While MOSFETs are more common in high-density circuits, bipolar transistors enable high-speed circuits because of their high transconductance. It therefore is obvious that MBE device activities should primarily be devoted to the

improvement of these devices. Another interesting topic is the optoelectronic behaviour of silicon-based devices.

6.5.1 Heterojunction Bipolar Transistors

The continuously variable energy gap of the SiGe system [40] allows one to optimize the various demands on bipolar transistors such as high current gain and high Early voltage.

The properties of Si/Si_{1-x}Ge_x/Si heterojunction bipolar transistors (HBTs) with a strained Si_{1-x}Ge_x base grown on (100) Si substrates were first reported in 1988 by Tatsumi et al. [41]. Several publications by various groups followed during the same year (for a review see [42]). The lower band gap in the base allows a better current gain of the HBT because of the increased collector current I_C as compared with a homojunction device (bipolar junction transistor, BJT). For a uniform Ge fraction in the base, the current enhancement can be approximated by

$$I_{C,HBT} / I_{C,BJT} \propto e^{\Delta W / kT} \quad (6.4)$$

where ΔW is the band gap reduction in the base.

In addition, a graded Si_{1-x}Ge_x profile according to Fig. 6.15 produces an additional electric field, which enhances the injection of electrons from the emitter into the collector, leading to shorter transit times and thus higher cut-off frequencies.

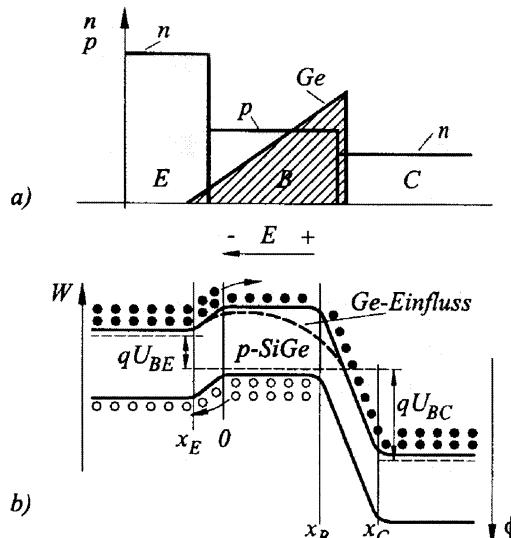


Fig. 6.15. Schematic illustration of (a) graded Ge concentration, and (b) band diagram of a heterojunction bipolar transistor [44]

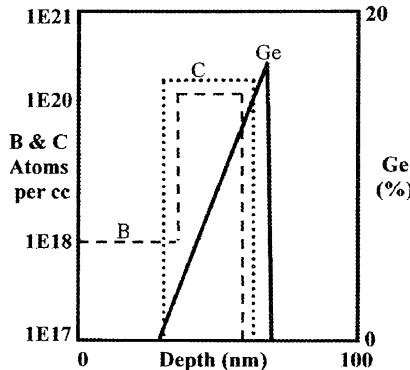


Fig. 6.16. Graded-Ge SiGe:C HBT [45]

As SiGe HBT base profiles become aggressively scaled to meet high performance requirements, the out-diffusion of B has become a severe problem. The addition of carbon to the B-doped regions of the base and slightly outside of this region on either side of the base (Fig. 6.16; see Chap. 8) can reduce this effect significantly, providing a narrower base region and thus even higher frequencies.

Despite the fact that most of the devices nowadays are fabricated with CVD techniques for production, most of the important issues have been investigated with MBE. As a result, from 1994 to today the maximum reported unity-current-gain cut-off frequency f_T has increased from 113 GHz to 170 GHz, and f_{max} has increased from 90 GHz to 350 GHz [43].

6.5.2 SiGe MOSFETs and MODFETs

One way to improve the electron and hole mobilities in a Si or SiGe channel is to stretch the lattice in order to achieve stress-induced valley splitting of the subband systems [46].

Owing to the fact that the hole mobility in silicon is smaller than the electron mobility, an NMOSFET always is faster than the complementary PMOSFET. From this one can deduce that the maximum speed of a CMOS inverter structure is limited by the PMOSFET. The idea behind one type of SiGe MOSFET is therefore to increase the hole mobility in a p-type channel. This can be achieved if the charge carriers are confined within a pseudomorphically strained $\text{Si}_{1-x}\text{Ge}_x$ channel, which can be grown by MBE. This offers the possibility to create faster CMOS inverters without changing the technology node.

However, the concept suffers from the problem that it is impossible to establish a high-quality gate oxide, which is needed for good high-frequency performance. The only known appropriate gate insulator material so far is SiO_2 thermally grown at temperatures above 750°C. But a high thermal

budget induces stress relaxation by introducing dislocations (lattice damage), causing a breakdown of the mobility improvement. Furthermore, Si and Ge start to interdiffuse, causing a smear-out of the multilayer heterostructure composition. This problem is solved by using a Si cap layer of a few nm which then can be oxidized by a low-temperature oxide.

For a SiGe NMOSFET, a strained Si channel has to be grown on a fully relaxed $\text{Si}_{1-x}\text{Ge}_x$ alloy. In Fig. 6.17, the classical channel structure is shown [47]. The source, channel and drain of a SiGe MOSFET are arranged in the same way as in a classical MOSFET.



Fig. 6.17. Schematic view of a SiGe MOSFET channel, rotated by 90° clockwise

To introduce the mechanical stress into the Si channel, a graded SiGe buffer is grown by means of MBE on a Si substrate, with a final relaxed, doped SiGe top layer with a constant Ge mole fraction used as a SiGe pseudo-substrate. On this pseudo-substrate, the doped, strained Si channel is deposited, also by means of MBE. For all further fabrication steps, standard lateral MOSFET technology is used.

In Fig. 6.18, the classical channel structure of an n-type SiGe MODFET is shown. The source, channel and drain of a SiGe MODFET are arranged in the same way as in a classical MOSFET.

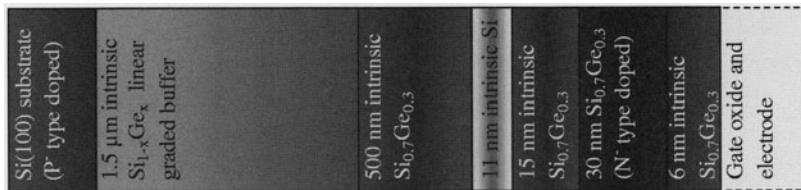


Fig. 6.18. Schematic view of a SiGe-based MODFET, channel rotated by 90° clockwise

With the gate electrode, one can control the electron concentration (dominated by the n⁻-type-doped SiGe layer) in the pseudomorphically stretched intrinsic Si layer with a thickness of a few nm. Owing to the lattice mismatch

between the $\text{Si}_{1-x}\text{Ge}_x$ layers and the Si layer, a mechanical stress is always present in the channel.

The electron mobility in this channel is much higher than the electron mobility in the inversion layer of a classical Si MOSFET on the Si channel/gate oxide interface. The reasons for this are: (1) Owing to the tensile stress, the intrinsic Si layer is $\text{Si}_{1-x}\text{Ge}_x$ -like (in general, the mobility in $\text{Si}_{1-x}\text{Ge}_x$ is higher than in pure Si), (2) no electron scattering on ionized scattering centres (ionized acceptors or donors) occurs, and (3) no electron interface scattering occurs.

For SiGe-based MODFETs, the fabrication of the MOS gate stack is also extremely difficult owing to the fact that Ge does not form a stable thermal oxide. It is hoped that the ongoing development of so-called “high- k ” dielectrics deposited by means of low-temperature LPCVD or ALD will yield solutions to this problem.

6.5.3 Vertical MOSFET Structures

As mentioned above, the main advantage of using MBE is the possible fabrication of ultra-sharp δ -profiles and the thickness control on an atomic scale of the deposited Si or SiGe films. This was the main reason why many researchers worldwide have used this technique to fabricate vertical devices with nanometre dimensions, independent of the advanced lithography method.

In 1993, for the first time, a fully MBE-fabricated vertical MOS device was presented. This new device concept was developed by Gossner et al. and was called a “planar-doped-barrier MOSFET” (PDBFET) [48]. In Fig. 6.19, the schematic structure of a PDBFET is shown.

The whole channel doping concentration is confined in a very sharp δ -doping profile. After all process steps, the effective δ -thickness is in the range of 6 nm (the as-grown thickness amounts to approximately 3 nm).

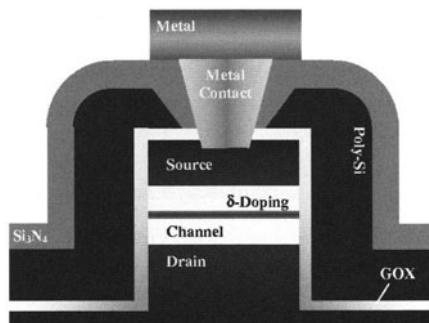


Fig. 6.19. Schematic structure of a vertical planar-doped-barrier MOSFET (PDBFET)

This δ -doping structure is flanked by lightly doped source and drain extensions (LDDs) and by the highly doped source and drain regions themselves. The doping concentrations inside the LDDs and inside the source and drain amount to $1 \times 10^{16} \text{ cm}^{-3}$ and $5 \times 10^{19} \text{ cm}^{-3}$, respectively. In their work, the authors of [48] investigated mainly the influence of LDD length and of δ -doping concentration.

From the graphs in Fig. 6.20 one can see that the drain-induced barrier lowering (DIBL) depends strongly on the δ -doping concentration. With a shorter distance between the source and drain, a higher δ -doping is necessary for a successful suppression of the DIBL. Here one well-known problem arises: with increasing δ -doping for DIBL suppression, one will lose V_T control because higher δ -doping causes a higher threshold voltage.

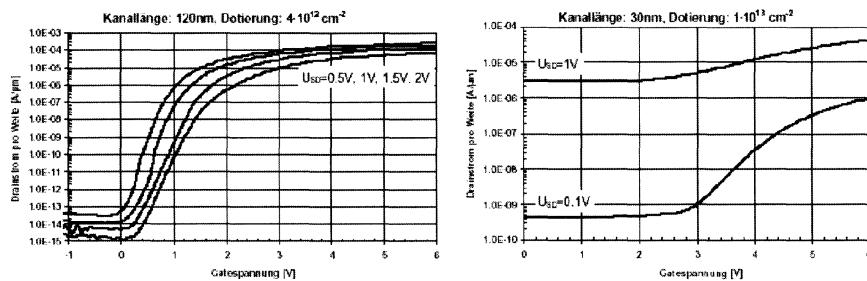


Fig. 6.20. Transfer characteristics of a PDBFET with a distance of 120 nm (left) and 30 nm (right) between source and drain. In both transistors the position of the δ -doping is symmetric and all doping levels are nearly equal [49]

At this point one retrospective remark should be made: at the panel discussion of the International Electron Devices Meeting (IEDM) 2001, the feasibility of a 10 nm bulk Si MOSFET was discussed and contradictions were apparent. Unfortunately, Gossner et al. incorrectly identified the distance between the source and drain as the channel length of their devices. However, the true channel length is given by the effective δ -thickness. Therefore, one can state that a positive answer to this question was given in 1993.

One major disadvantage of all vertical device structures fabricated by means of MBE is the fact that the gate structure cannot be fabricated so that it is self-aligned with the source and drain. Therefore a vertical MOSFET is heavily loaded with parasitic overlap capacitances. This is the reason why high-speed electronics for logic applications with high clocking frequencies are still realized with lateral MOSFET technology.

Therefore, in MBE research for manufacturing of vertical devices, one focus was placed on vertical device structures for power electronics, where parasitic overlap capacitances are not a “killing argument”.

A vertical power MOSFET is characterized by a lightly doped (or undoped) drift zone between the channel region and the drain acting as a bleeder. The breakdown voltage in the off-state of the power transistor is defined by the length of this drift zone (several microns). Two main contributions to the on-state resistance R_{ON} – which determines the power losses during operation – can be identified: (1) the channel region, and (2) the drift zone. Fink showed that the series resistance of the channel region can be more effectively reduced if the PDBFET concept is used instead of vertical power concepts with homogeneously doped channel regions [50]. From Fig. 6.21, one can clearly see that for identical gate voltages the on-current between the source and drain increases significantly for a power PDBFET.

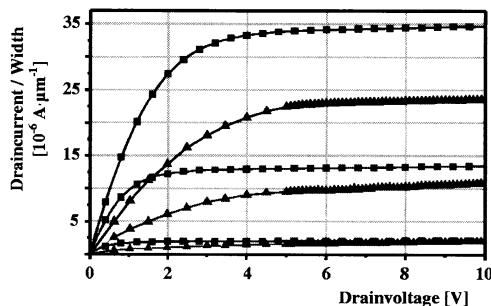


Fig. 6.21. Output characteristics of a vertical power PDBFET (squares) compared with vertical power MOSFET with a homogeneously doped channel region (triangles) with identical channel length. $V_G - V_T = 1, 2, 3$ V [50]

6.6 Selected Research Highlights

6.6.1 Cascade Laser

Quantum cascade (QC) structures offer a viable method for the fabrication of Si-based lasers, because these devices rely on intersubband transitions only. These are so-called direct transitions, with a high quantum efficiency for the generation of light as compared with the indirect transitions between the conduction and valence bands. Using Si/SiGe quantum wells, the energetic subband splitting allows – depending on composition – a wavelength regime between 3 and 300 μ m. In order to achieve a high efficiency, multiple quantum wells in the injector and collector regions with a fully developed miniband structure are grown by MBE. A schematic drawing of the calculated valence band structure for a strain-compensated QC superlattice is depicted in Fig. 6.22 [51].

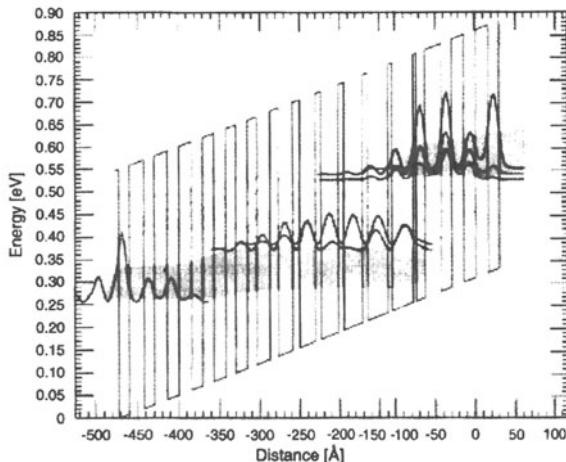


Fig. 6.22. Valence band edge of Si/SiGe quantum cascade structure

Besides interband photoluminescence measurements, the electroluminescence has also been investigated. Pronounced peaks at 80 K have been observed at an energy of 176 meV, which agrees reasonably well with the expected value. These results are one important step towards a Si/SiGe-based laser.

6.6.2 Resurf Structures

If the channel length of a power MOSFET is small compared with its drift zone, and if the drift zone length itself is several tens of microns, then the series resistance of the drift zone is the dominant part of the total on-resistance R_{ON} , which should be as low as possible. On the other hand, in the off-state the drift zone determines the breakdown voltage. The highest breakdown voltages can be achieved if the drift zone is ideally intrinsic, i.e. R_{OFF} must be as high as possible. One way out of this dilemma is a dynamic resistance, which can be realized by so-called resurf (reduced surface field) structures.

One possible lateral multi-resurf structure is the well-known $N(\delta ni\delta pi)$ structure fabricated by means of MBE [52–54]. As shown in Fig. 6.23, the base ($\delta ni\delta pi$) structure is an alternating layer sequence with an n-type δ -doping followed by an undoped (intrinsic) layer, a p-type δ -doping, and again an intrinsic layer. N gives the number of repetitions of ($\delta ni\delta pi$).

As shown by Döhler in 1982, such $N(\delta ni\delta pi)$ structures can be designed to be fully compensated [55]. This means that all acceptors and all donors in the structure are ionized (the donor electrons occupy the acceptor states) and the number of free electrons and holes is given only by the intrinsic carrier concentration of Si, i.e. a well-designed $N(\delta ni\delta pi)$ structure acts like intrinsic Si.

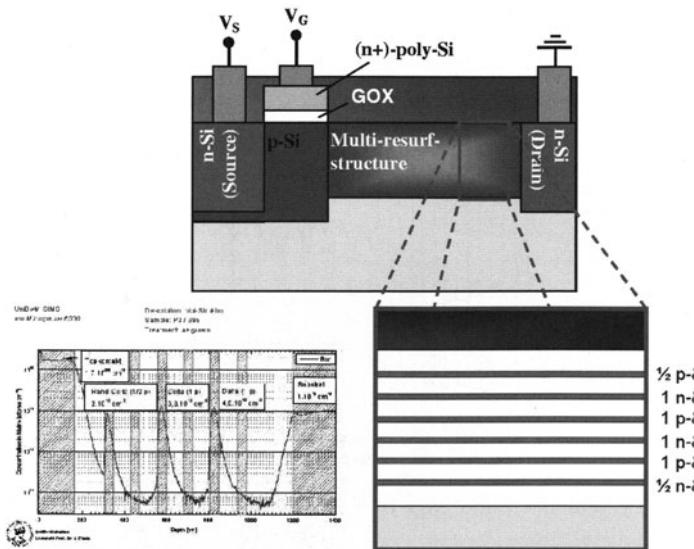


Fig. 6.23. Schematic structure of a lateral power MOSFET with a multi-resurf structure used as a drift zone (*top*); this multi-resurf structure can be designed as a $(\delta n i \delta p i)$ structure (*bottom right*); SIMS profile of the p-type doping in this $(\delta n i \delta p i)$ structure (*bottom left*) [52–54]

This explains why $N(\delta n i \delta p i)$ structures are interesting for high-power electronics. In the off-state of the transistor, the $N(\delta n i \delta p i)$ drift zone is fully compensated and behaves like intrinsic Si. In this case the breakdown voltage is determined only by the drift zone length and the critical electrical field for breakdown of intrinsic Si. In the on-state, electrons coming from the source are injected into the drift zone and flow to the drain in the highly n-type doped layers of the $N(\delta n i \delta p i)$ drift zone with a small series resistance.

6.6.3 Self-Organization and Ordering

One of the major topics in Si and SiGe MBE research is the investigation of self-organized growth and ordering phenomena. This provides a possible alternative method for implementing optoelectronic and nanoelectronic applications without the drawback of using advanced lithography, thus avoiding process-induced damage. The potential device applications of self-organized nanostructures can be found in review papers [56, 57]. In the following subsections, prominent areas of this research will be presented.

Self-Organization of Si and SiGe by Micro-Shadow Mask Technique

The micro-shadow mask technique for self-organized Si growth was first published in 1993 by Hammerl and Eisele [58]. Two years later, Brunner et al. discussed self-organized SiGe growth [59]. Self-organized growth by using micro-shadow masks is a unique molecular-beam epitaxy technique. Initially, self-organized growth of Si and SiGe stripes was under investigation (see Fig. 6.24).

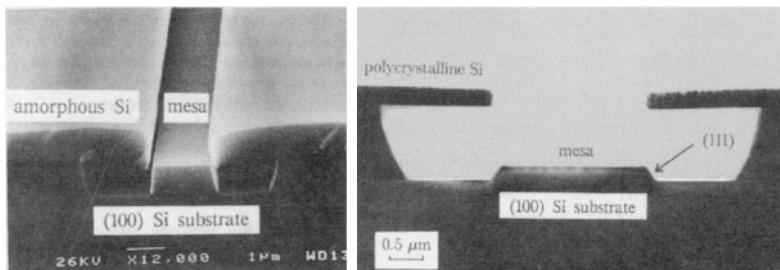


Fig. 6.24. SEM micrograph of locally grown and cleaved Si mesa stripes (*left*) and cross-sectional TEM micrograph of a locally grown Si mesa structure with $\{111\}$ faceted mesa side-walls (*right*) [58]

Self-organized growth as a function of deposition temperature, growth rate, mask aperture and mask alignment has been investigated. Gossner et al., for instance, investigated the self-organized formation of Si pyramids with four $\{111\}$ faceted side-walls on a Si (100) substrate as a function of mask alignment with respect to the substrate [60]. They showed that only in an initial growth stage do the Si mesa islands correspond to the mask shape and alignment. If growth continues, the influence of the mask is negligible and pyramidal structures of identical size and shape are formed (see Fig. 6.25).

These results could be explained by a model which is based on the experimental fact that the $\{111\}$ Si facet is one of the most stable Si facets. In terms of energy, this means that the surface free-energy function has a minimum for the $\{111\}$ facet and the driving force for self-organized growth is the reduction of the total surface free energy [61].

Germanium Quantum Dots on Silicon

In the long list of publications on Si and SiGe MBE research, the formation of Ge quantum dots – also known as “hut-clusters” – has a leading position. For self-organized Ge dots on Si (100) substrates, one of the main advantages is compatibility with the present Si ULSI technology.

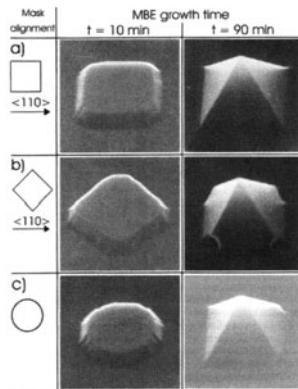


Fig. 6.25. SEM micrographs of Si mesa structures grown by MBE with micro-shadow masks, in growth states after 10 and 90 min. The mask alignment with respect to the crystallographic orientation and the shape of the mask aperture are indicated. The mask opening was 250 nm and the growth temperature was kept constant at 500°C [60]

The strain caused by the 4.2% difference in the lattice parameters of Si and Ge can lead to island formation in the Stranski-Krastanov mode [62]. However, owing to the random process of island nucleation, the self-assembly technique suffers from broad distributions in both size and position [63]. In recent years it has been shown that, starting from a single layer with inhomogeneous islands, one can greatly improve the size homogeneity by growing several multilayer structures, i.e. structures in which the layers containing Ge islands are separated by Si spacer layers. In this case the islands in the upper layers grow on top of the buried islands, giving rise to a vertical correlation between islands along the growth direction. A cross-sectional TEM image of a sample with 10 Ge layers (nominally 4 monolayers thick) separated by 22 nm Si spacers is shown in Fig. 6.26 [65]. The image clearly shows that each island in the upper layers grows on top of the islands in the lower layers.

Besides the size homogeneity, the spatial distribution of the dots is important if one wishes to exploit dot arrays for computational and signal-processing applications, for example, quantum-dot cellular automata [66]. One of the most interesting approaches for the arrangement of self-organized Ge dots is to use selective epitaxial growth (SEG) of Si mesas (prepared by micro-shadow mask technique) as templates for subsequent Ge growth.

As can be seen from Fig. 6.27, one-dimensional dot arrays have been grown by MBE on the ridges of Si mesas [67]. The average spacing may be attributed to the balance between the strain energy of the dots and the repulsive interactions of neighbouring dots through the substrate.

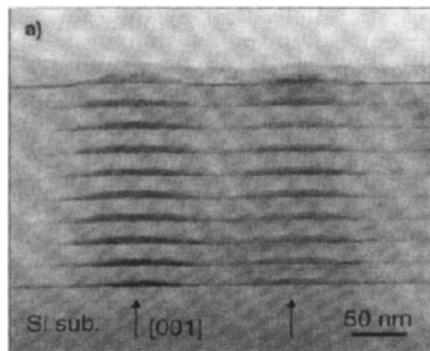


Fig. 6.26. Cross-sectional TEM image taken along the [011] azimuth of a sample containing 10 Ge/Si bilayers [65]

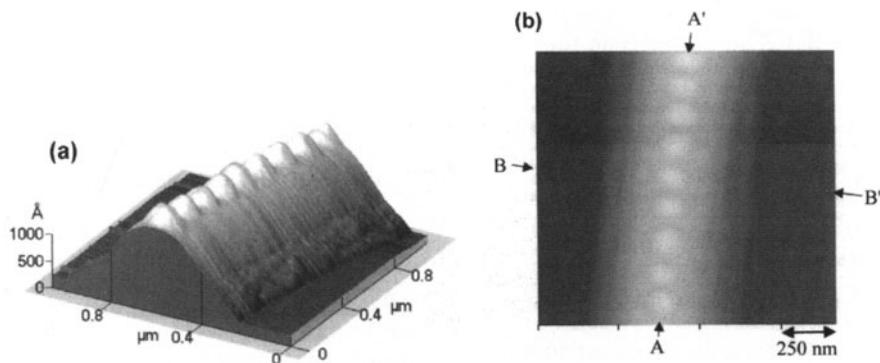


Fig. 6.27. Self-aligned one-dimensional array of Ge dots (10 monolayers Ge) on the ridge of a Si mesa stripe oriented in the (110) direction: (a) 3D AFM image, and (b) 2D image of the sample as seen from the top [67]

Self-Organization of Boron on Silicon

To reach the ideal case of a δ -layer with a thickness of just one atomic layer, Zotov et al. [68] and Baumgärtner et al. [69] proposed the use of the “ordered δ -doping technique”, i.e., the formation of surface phases with doping atoms ordered periodically in a 2D plane.

Surface phases on Si represent 2D superlattices on top of a Si substrate. Many adsorbate materials and their surface phases on different Si orientations have been examined and studied in recent years [70]. Regarding B, which is one of the important doping materials, the most interesting surface phase is the $T_4-\sqrt{3} \times \sqrt{3}-R30^\circ$ -B surface phase (T_4 -BSP) on top of a Si (111) substrate [71, 72, 74, 75]. By MBE techniques, B is deposited onto the Si surface at a substrate temperature near 600°C with a deposition of 1/3 ML (1/3 monolayer = 2.6×10^{14} B atoms per cm^2). In this phase one B atom

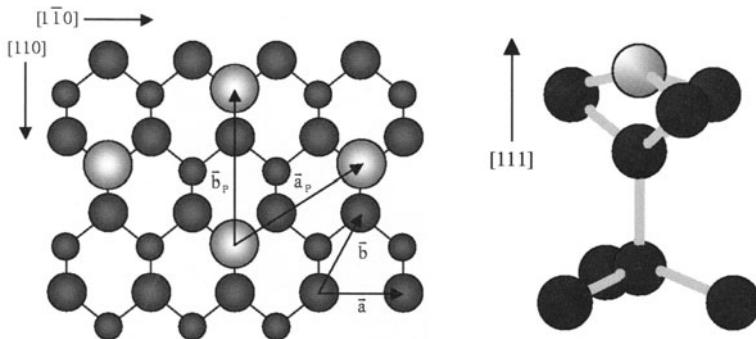


Fig. 6.28. B atoms (light circles) on top of a Si (111) surface (dark circles). Each B atom saturates three Si top-layer atoms of the Si (111) double layer. The primitive cell of the 2D Si surface lattice (vectors \mathbf{a} and \mathbf{b}) and the primitive cell of the T₄-BSP with an area of $A = 0.383 \text{ nm}^2$, are also shown (vectors $\mathbf{a}_P = \sqrt{3}\mathbf{a}$ and $\mathbf{b}_P = \sqrt{3}\mathbf{b}$). The B density in the surface phase $2.6 \times 10^{14} \text{ cm}^{-2} \equiv 1/3 \text{ ML}$

resides in a T₄ site atop the Si (111) plane saturating three Si dangling bonds (see Fig. 6.28) [72].

STM studies by Stimpel et al. (see Fig. 6.29) show a three-step formation process for the T₄-BSP, with a 7 × 7 reconstructed Si (111) surface as the starting point [76].

In the first step the B atoms, deposited at a temperature near 600°C, start to substitute for the Si top atoms in the subunits of the 7 × 7 reconstruction. After replacement of around 50% of all Si top atoms of one 7 × 7 unit cell by B,

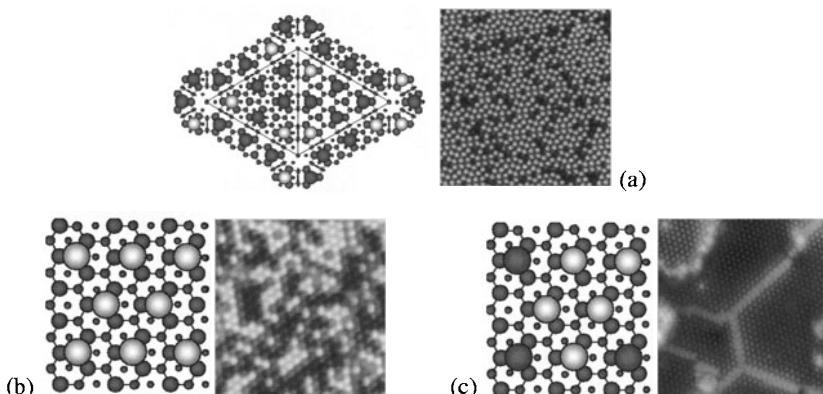


Fig. 6.29. (a) B atoms (light circles in the schematic picture) substitute for Si top atoms in the subunits of the 7 × 7 reconstructed Si (111) surface (STM image: $A = 17 \times 17 \mu\text{m}^2$); (b) breakdown of the 7 × 7 reconstruction: Si and B atoms together form a T₄- $\sqrt{3} \times \sqrt{3}$ -R30° reconstruction (STM image: $A = 17 \times 17 \text{ nm}^2$); (c) a complete T₄-BSP is formed (STM image: $A = 17 \times 17 \text{ nm}^2$)

the 7×7 reconstruction breaks down, and together the Si and B atoms form a $T_4-\sqrt{3} \times \sqrt{3}-R30^\circ$ reconstruction (second step). A further B deposition leads to the displacement of the remaining Si atoms in the $T_4-\sqrt{3} \times \sqrt{3}-R30^\circ$ reconstruction (third step), and after the deposition of $1/3$ ML of B a complete T_4 -BSP is formed.

Further STM studies revealed that the T_4 -BSP is stable for a temperature stress range $< 825^\circ\text{C}$. At higher temperatures the boron atoms start to diffuse by two-step diffusion via the S_5 crystal lattice positions into the silicon substrate.

It is interesting to note that the T_4 -BSP and also the T_4 -SiSP in combination with the S_5 -BSP can cause self-organized growth. The fully passivated surface allows a large surface diffusion length, which leads to nucleation of impinging atoms and thus to nanodots. Schulze et al. observed self-organized MBE growth of triangular pyramidal $\{113\}$ faceted germanium quantum dots at 400°C on the T_4 -BSP. The dots are uniform in mechanical stress and size, and aligned in parallel chains (see Fig. 6.30) [64, 77].

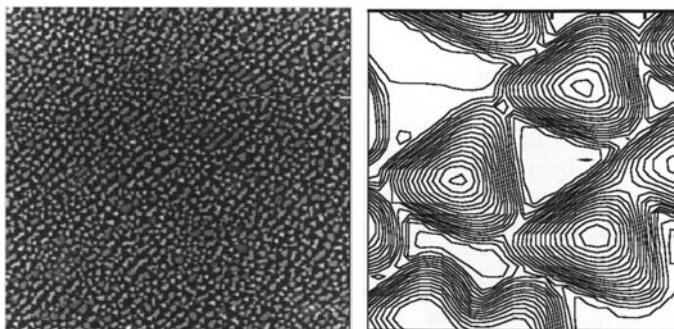


Fig. 6.30. Ge quantum dots on the T_4 -BSP on (111) Si: AFM picture ($A = 15 \times 15 \mu\text{m}^2$) (left) and contour lines ($A = 0.7 \times 0.7 \mu\text{m}^2$) (right)

6.7 Conclusion

During the past decades, Si MBE has been an extremely valuable tool for the investigation of novel layer structures with thickness control on an atomic scale. The successful growth of abrupt and δ -type doping profiles and of SiGe heterostructures has contributed significantly to the understanding of new materials and devices for microelectronics. The ongoing research activities are directed towards self-limiting and self-organizing processes which will be necessary in order to fulfil the requirements set out in the ITRS roadmap for scaling devices to nanometer dimensions.

References

1. B.A. Unvala: *Nature* **194**, 966 (1962)
2. A.P. Hale: *Vacuum* **13**, 93 (1963)
3. H.C. Abbink, R.M. Broudy, G.P. McCarthy: *J. Appl. Phys.* **10**, 4673 (1968)
4. R.N. Thomas, M. Francombe: *Surf. Sci.* **25**, 357 (1971)
5. F. Jona: *Appl. Phys. Lett.* **9**, 235 (1966)
6. G. Booker, B. Joyce, R. Bradley: *Phil. Mag.* **10**, 1087 (1964)
7. E. Handelmann, E. Povilonis: *J. Electrochem. Soc.* **111**, 201 (1964)
8. V.E. Kuznetsov, V. Postnikov: *Sov. Phys. Cryst.* **16**, 357 (1969)
9. Y. Nannichi: *Nature* **200**, 1087 (1963)
10. D. Thomas: *Phys. Stat. Sol.* **13**, 359 (1966)
11. L. Weisberg: *J. Appl. Phys.* **38**, 4537 (1967)
12. H. Widmer: *Appl. Phys. Lett.* **5**, 108 (1964)
13. E. Kasper, K. Werner: *J. Electrochem. Soc.* **132**, 2481 (1985)
14. E. Kasper, J.C. Bean. In: *Silicon Molecular Beam Epitaxy I, II* (CRC Press, Boca Raton 1988)
15. E. Kasper, H.J. Herzog, H. Kibbel: *Appl. Phys.* **8**, 199 (1975)
16. E. Kasper, H.J. Herzog: *Thin Solid Films* **44**, 357 (1970)
17. J.C. Bean, L.C. Feldmann, T. Sheng, A. Fiory, R. Lynch: *J. Vac. Sci. Technol. A* **2**, 436 (1984)
18. E. Kasper, A. Schuh, G. Bauer, B. Hollander, H. Kibbel: *J. Cryst. Growth* **157**, 68 (1995)
19. J.P. Dismukes, L. Ekstron, R.J. Paff: *J. Phys. Chem.* **68**, 3021 (1964)
20. E. Kasper, K. Lyutovich (eds.): EMIS Datareview 24: *Properties of SiGe:C* (IEE, London 2000)
21. H.J. Osten, G. Lippert, P. Gaworzewski, R. Sorge: *Appl. Phys. Lett.* **71**, 1522 (1997)
22. J. Murata, A. Moriya, M. Sakuraba: Proc. 8th Int. Symp. Si Mater. Sci. Technol., San Diego (1998) p. 822
23. G.E. Becker, J.C. Bean: *J. Appl. Phys.* **48**, 3395 (1977)
24. U. König, H. Kibbel, E. Kasper: *J. Vac. Sci. Technol.* **16**, 985 (1979)
25. E. Friess, J. Nützel, G. Abstreiter: *Appl. Phys. Lett.* **60**, 2237 (1992)
26. H.J. Gossmann. In: *Delta-Doping of Semiconductors* (Cambridge University Press 1996) p. 161
27. H. Jorke: *Surf. Sci.* **193**, 569 (1988)
28. H.J. Gossmann, F.C. Unterwald: *Phys. Rev. B* **47**, 12618 (1993)
29. J.A. Venables, G.L. Price. In: *Epitaxial Growth* (Academic Press 1975) p. 381
30. D.J. Eaglesham: *J. Appl. Phys.* **77**, 3597 (1995)
31. J. Störmer, P. Wilutzki, D.T. Britton, G. Kögel, W. Triftshäuser, W. Kiunke, F. Wittmann, I. Eisele: *Appl. Phys. A* **61**, 71 (1995)
32. T. Rupp: Ph.D. Thesis, Universität der Bundeswehr München (1996)
33. H.P. Zeindl, T. Wegehaupt, I. Eisele, H. Oppolzer, H. Reisinger, G. Tempel, F. Koch: *Appl. Phys. Lett.* **50**, 1165 (1987)
34. I. Eisele. In: *Delta-Doping of Semiconductors* (Cambridge University Press 1996) p. 137
35. H.P. Zeindl, T. Wegehaupt, I. Eisele: *Thin Solid Films* **184**, 21 (1990)
36. N.L. Matthey, M.G. Dowsett, E.H.C. Parker, T.E. Whall, S. Taylor, J.F. Zhang: *Appl. Phys. Lett.* **57**, 1648 (1990)

37. H. Jorke, H.J. Herzog, H. Kibbel: *Appl. Phys. Lett.* **47**, 511 (1985)
38. H.P. Zeindl, E. Hammerl, W. Kiunke, I. Eisele: *J. Electron. Mater.* **19**, 1119 (1990)
39. I. Eisele: *Superlatt. Microstruct.* **6**, 123 (1989)
40. C. Penn, T. Fromherz, G. Bauer. In: *Silicon Germanium and SiGe:Carbon*, ed. by E. Kasper, K. Lyutovich (INSPEC, London 2000) p. 125
41. T. Tatsumi, H. Hirayama, N. Aizaki: *Appl. Phys. Lett.* **52**, 895 (1988)
42. J.C. Sturm, H. Yin. In: *Silicon Germanium and SiGe:Carbon*, ed. by E. Kasper, K. Lyutovich (INSPEC, London 2000) p. 305
43. J.S. Rieh et al.: *IEDM Technical Digest* (2002) p. 771
44. K. Hoffmann. In: *Systemintegration* (Oldenbourg Verlag 2003)
45. D. Meyer, B. Pagliaro, D. Webb, J. Sellar, M. Ward, J. Robinson, S. Young: *Proc. 3rd Int. Conf. SiGe(C) Epitaxy and Heterostructures*, Santa Fe (2003) p. 53
46. J.L. Hoyt, H.M. Nayfeh, S. Eguchi, I. Aberg, G. Xia, T. Drake, E.A. Fitzgerald, D.A. Antoniadis: *Tech. Dig. Int. Electron Devices Meet.*, San Francisco (2002) 2.1.1
47. J. Welser, J.L. Hoyt, J.F. Gibbons: *Tech. Dig. Int. Electron Devices Meet.*, San Francisco (1992) 31.7
48. H. Gossner, W. Kiunke, I. Eisele, L. Risch, K. Hofmann, R. Treichler, H. Cerva: *Proc. Int. Conf. Solid State Devices Mater.*, Chiba (1993) PB-1-20
49. F. Kaesen: Ph.D. Thesis, Universität der Bundeswehr München (1998)
50. C. Fink: Ph.D. Thesis, Universität der Bundeswehr München (2000)
51. D. Grützmacher et al.: *J. Cryst. Growth* **251**, 707 (2003)
52. C. Tolksdorf: *Proc. 60th Annual Device Res. Conf.*, Santa Barbara (2002) p. 85
53. C. Tolksdorf: *Proc. 61st Annual Device Res. Conf.*, Salt Lake City (2003) p. 59
54. C. Tolksdorf: Ph.D. Thesis, Universität der Bundeswehr München (2003)
55. G.H. Döhler: *Proc. Int. Conf. Solid State Devices Mater.*, Tokyo (1982) A-1-4
56. M.J. Kelly: *Semicond. Sci. Technol.* **5**, 1209 (1990)
57. H. Luth: *Appl. Surf. Sci.* **130-132**, 855 (1998)
58. E. Hammerl, I. Eisele: *Appl. Phys. Lett.* **62**, 2221 (1993)
59. J. Brunner, W. Jung, P. Schnittenhelm, M. Gail, G. Abstreiter, J. Gonderman, B. Hadam, T. Koester, B. Spangenberg, H.G. Roskos, H. Kurz, H. Gossner, I. Eisele: *J. Cryst. Growth* **157**, 270 (1995)
60. H. Gossner, T. Rupp, I. Eisele: *J. Cryst. Growth* **157**, 308 (1995)
61. H. Baumgärtner, F. Kaesen, H. Gossner, I. Eisele: *Appl. Surf. Sci.* **130-132**, 747 (1998)
62. J. Tersoff, Y.H. Phang, Z. Zhang, M.G. Lagally: *Phys. Rev. Lett.* **75**, 2730 (1995)
63. V. Le Thanh, P. Boucaud, D. Debarre, Y. Zheng, D. Bouchier, J.M. Lourtioz: *Phys. Rev. B* **58**, 13115 (1998)
64. J. Schulze: Ph.D. Thesis, Universität der Bundeswehr München (2000)
65. V. Le Thanh, V. Yam, P. Boucaud, Y. Zheng, D. Bouchier: *Thin Solid Films* **369**, 43 (2000)
66. A.O. Orlov, I. Amlani, G.H. Berstein, C.S. Lent, G.L. Snider: *Science* **277**, 928 (1997)
67. G. Jin, J.L. Liu, Y.H. Luo, K.L. Wang: *Thin Solid Films* **369**, 49 (2000)
68. A.V. Zотов, V.G. Lifshits, T. Rupp, I. Eisele: *J. Appl. Phys.* **83**, 5865 (1998)
69. H. Baumgärtner, W. Hansch, F. Wittmann, I. Eisele: *Curr. Topics Cryst. Growth Res.* **2**, 283 (1995)

70. V.G. Lifshits, A.A. Saranin, A.V. Zotov: *Surface Phases on Si: Preparation, Structures, and Properties* (Wiley 1994)
71. Z. Zhang, M.A. Kulakov, B. Bullemer, I. Eisele: *J. Vac. Sci. Technol. B* **14**(4), 2684 (1996)
72. A.V. Zotov, S.V. Ryzhkov, V.G. Lifshits: *Surf. Sci.* **328**, 95 (1995)
73. D.J. Tweet, K. Akimoto, T. Tatsumi, I. Hirosawa, J. Mizuki, J. Matsui: *Phys. Rev. Lett.* **15**, 2236 (1992)
74. R.L. Headrick, B.E. Weir, A.F.J. Levi, D.J. Eaglesham, L.C. Feldman: *J. Cryst. Growth* **111**, 838 (1991)
75. P. Avouris, I.W. Lyo, F. Bozso, E. Kaxiras: *J. Vac. Sci. Technol. A* **8**(4), 3405 (1990)
76. T. Stimpel, J. Schulze, H.E. Hoster, I. Eisele, H. Baumgärtner: *Appl. Surf. Sci.* **162-163**, 382 (2000)
77. J. Schulze, H. Baumgärtner, C. Fink, G. Dollinger, I. Genchev, L. Görgens, W. Hansch, H.E. Hoster, T.H. Metzger, R. Paniago, T. Stimpel, T. Sulima, I. Eisele: *Thin Solid Films* **369**, 10 (2000)

7 Amorphous Hydrogenated Silicon, a-Si:H

W. Fuhs

7.1 Introduction

The research in the field of amorphous semiconductors, from the very beginning, has been driven by both the scientific interest in basic aspects of disorder in the properties of solids and technological applications. In the early years chalcogenide glasses were at the center of the interest owing to thin-film applications in imaging, xerography, memory and switching devices. At that time amorphous silicon and amorphous germanium, a-Si and a-Ge, were of more academic scientific interest. As simple elemental tetrahedrally bonded amorphous semiconductors, they served as model systems in which the disorder was less complicated, being defined not by chemical composition but by the structural disorder only. These amorphous semiconductors were prepared as thin films, about $0.1\text{--}1\,\mu\text{m}$ thick, on glass or quartz substrates by a variety of methods such as thermal or e-gun evaporation, sputtering, ion bombardment and electrolysis. The simplest model for the structure of tetrahedrally bonded semiconductors is the continuous random network (CRN) structure in which the average coordination number is close to 4. Fluctuations in the bond angles and nearest-neighbor distances lead to a loss of long-range order even in the second-neighbor shell. This loss of long-range order is the characteristic structural feature of amorphous semiconductors. As a result, important theoretical concepts which are based on periodicity (Bloch's principle) fail, such as band structure, k -vector, Bloch states, effective masses and optical selection rules. The optical spectra of amorphous semiconductors appear to be more or less broadened versions of their crystalline counterparts, which shows that the density-of-states distribution is the decisive quantity; this is largely determined by the nature and structure of the chemical bonding. Perhaps the most obvious effect of disorder is the localization of electronic states, in particular near the band edges, which strongly affects the transport properties.

In the 1960s the major challenge consisted in understanding the role of disorder and in developing new theoretical concepts. However, the early forms of amorphous silicon had unacceptable electronic properties due to large defect densities, which caused high densities of states in the energy gap, which pinned the Fermi level. The only defect that has been identified microscopically by electron spin resonance is the Si dangling bond (Si-db), which is present in an amount of some $10^{19}\,\text{cm}^{-3}$ in evaporated or sputtered material.

Si dangling bonds form deep gap states, which act as effective centers for nonradiative recombination. The conductivity in this kind of amorphous silicon was very low and, below room temperature, was in general determined by hopping transport between localized gap states. Owing to the effective pinning of the Fermi level in a high density of gap states, the conductivity could not be varied by doping, illumination or carrier injection. Such material properties prevented this material from being useful for electronic devices.

The situation changed rather abruptly when the beneficial role of hydrogen incorporation was discovered at the beginning of the 1970s. Six main milestones in the material and device research led to a burst of research activities in this field and to numerous applications of hydrogenated amorphous silicon, a-Si:H:

- The first to use the plasma deposition technique were Chittick, Alexander and Sterling in 1969 [1]. In this technique, silane (SiH_4) was decomposed in a radio-frequency (rf) glow discharge and the film was formed on heated substrates.
- The work at the University of Dundee showed that the defect density in this kind of amorphous silicon was low, which resulted in high photoconductivity [2].
- The essential role of hydrogen in passivating defects was first discovered by the Harvard group, studying sputtered a-Si:H and a-Ge:H [3]. Later on, numerous studies gave proof that the superior semiconducting and photoelectric properties of glow-discharge-deposited amorphous silicon were due to the incorporation of hydrogen.
- In 1975 Spear and LeComber [4] reported on substitutional n-type and p-type doping by addition of phosphine or diborane to the process gas. This procedure allowed one to control the electrical conductivity over 10 orders of magnitude.
- The first report on photovoltaic devices, by Carlson and Wronski in 1976 [5], demonstrated the feasibility of a-Si:H solar cells.
- The research on displays started some years later, after the first report on the fabrication and physics of a thin-film transistor [6].

The preparation and properties of hydrogenated amorphous silicon have been described in numerous review articles and monographs (e.g. [7--11]), and the development is well documented in the proceedings of the biannual International Conference on Amorphous and Microcrystalline Semiconductors, which are published in regular issues of the Journal of Non-Crystalline Solids. Today hydrogenated amorphous silicon, a-Si:H, offers a mature material and device technology, used for solar cells, thin-film transistors, sensors, imaging, radiation detectors and displays. Among the various material choices for thin-film solar cells, this is the only technology which so far has been able to overcome the barrier to mass production of large-area modules and to occupy a reasonable share of the world market (about 10%).

7.2 Preparation and Structural Properties of Amorphous Silicon

Amorphous silicon cannot be made by rapid cooling of a Si melt. The amorphous material, instead, is prepared by deposition from the gas phase onto substrates which are held at temperatures far below the melting temperature. A large variety of techniques has been used: thermal evaporation, sputtering, chemical vapor deposition using silane (CVD), photo-CVD, plasma-enhanced chemical vapor deposition (PECVD), and thermo-catalytic hot-wire deposition (HWCVD). There are no differences in principle in the microstructures of amorphous films prepared by the various methods. The differences lie in the deposition rate and the kind and concentration of defects (dangling bonds and voids). PECVD has led to the lowest defect densities and, therefore, is widely used now in industrial applications.

Thermodynamically, amorphous silicon (a-Si) is in a metastable state. Although an ideal structure might be described by a random network structure, the real structure can be varied experimentally in many ways. Therefore the material properties strongly depend on the preparation conditions and on the thermal history of an amorphous sample. Thermal annealing has been shown to produce changes in practically all material properties (enthalpy, electrical properties, defect densities, optical properties, etc.) [12]. It has been found that the free energy of annealed a-Si (relaxed state) is about 0.11 eV/atom higher than that of crystalline silicon [13]. Heating above a temperature of about 500°C usually induces a transition into the thermodynamically more stable crystalline phase. The kinetics of solid phase crystallization (SPC) are characterized by nucleation and growth of crystal clusters at the expense of the surrounding amorphous material. Laser recrystallization of amorphous silicon has become an important technique for the fabrication of polycrystalline silicon thin-film transistors. Recently, SPC has attracted high interest as a result of the experience that the presence of certain metals strongly modifies the nucleation and growth process. Such techniques enable the engineering of polycrystalline silicon thin films at low temperatures for new device applications such as thin-film transistors or polycrystalline silicon thin-film solar cells. A particularly interesting example of this kind of processing is the creation of a polycrystalline silicon layer with a grain size of about 10 µm on a glass substrate by an Al-induced layer exchange process [14]. In this process, a film stack of glass/Al/a-Si is transformed by annealing at about 400°C into glass/poly-Si/Al(Si) such that after etching of the Al(Si) layer a polycrystalline silicon layer on glass remains, which may serve as a seeding layer in subsequent deposition processes.

In PECVD, silane (SiH_4) or mixtures of silane with rare gases or hydrogen are decomposed in a glow discharge. In most cases, parallel-plate systems in a stainless steel reactor have been used. The most important deposition parameters are substrate temperature, base pressure, flow rate of the process gas, power density and frequency. High-quality a-Si:H films are grown at de-

position rates of typically 0.1 nm/s at substrate temperatures in the range $150\text{--}250^\circ\text{C}$. The progress achieved by using this method is that the density of deep defects (Si dangling bonds) can be reduced to values of below 10^{16} cm^{-3} owing to the incorporation of hydrogen. The effect of hydrogen is to saturate dangling-bond defects and to lower the average coordination number, which allows the construction of a more relaxed disordered network. The hydrogen content can vary widely with the deposition conditions (5–40%). Infrared spectroscopy reveals various bonding configurations of hydrogen. In optimized films, the hydrogen concentration amounts to 5–15% and the density of neutral Si dangling bonds is less than 10^{16} cm^{-3} . The IR spectra of such films show that the hydrogen is bonded predominantly in an isolated Si-H configuration (absorption at 2000 cm^{-3}). However, NMR studies suggest that the hydrogen may be inhomogeneously distributed. Annealing of the films at temperatures above the deposition temperature leads to the evolution of hydrogen, which is accompanied by an enhancement of the defect density. In films deposited at around 250°C , the evolution rate peaks at 550°C , whereas films made at lower temperatures tend to show a second evolution peak near 350°C . The occurrence of this low-temperature peak points to the existence of a porous structure, which enables rapid hydrogen diffusion. Owing to the pronounced role of hydrogen, this kind of amorphous silicon may be considered rather as a hydrogen–silicon alloy. Practically all film properties depend on the hydrogen content [7–10].

A particular advantage of plasma deposition is a high flexibility in the choice of the process gases, which allows one to easily modify the properties of the deposited films. Variations of the deposition parameters have been found to result in higher deposition rates and modifications of the material. Doping can be achieved by adding controlled amounts of B_2H_6 or PH_3 to the process gas. Alloy films such as $\text{a-Si}_{1-x}\text{Ge}_x:\text{H}$ and $\text{a-Si}_{1-x}\text{C}_x:\text{H}$ can be deposited from gas mixtures of SiH_4 with GeH_4 or CH_4 , respectively. These films allow one to tune the energy gap to lower or higher values by controlling the composition of the process gas. Strong hydrogen dilution of the process gas results in the formation of microcrystalline silicon films (μc-Si:H). The structural and electronic properties of such plasma-deposited films have been studied intensively over more than 20 years [7–11].

7.3 Electronic Properties of Hydrogenated Amorphous Silicon, a-Si:H

The electronic properties of the amorphous films depend sensitively on the density and energy distribution $N(E)$ of the localized gap states. In particular, these states determine the shape of the absorption edge, the doping efficiency, transport and recombination, as well as the widths of space charge layers in devices. Figure 7.1 displays schematically a model of $N(E)$, including both intrinsic and extrinsic defect states. Inside the bands, the density-of-

states (DOS) distributions of the valence and conduction bands of amorphous semiconductors differ only little from those of their crystalline counterparts. In the optical spectra, the main effect is a broadening of the spectra. The states inside the bands are considered to be delocalized but, of course, owing to the loss of long-range order, they are no longer Bloch states. An important effect of disorder is that closer to the band edges, the states become localized. The most widely applied model suggests that a transition from extended to localized states occurs at distinct energies E_C and E_V , where the carrier mobilities drop abruptly owing to the change in the character of the states [7, 11]. In an amorphous semiconductor, these mobility edges play a similar role to the band edges in crystalline materials.

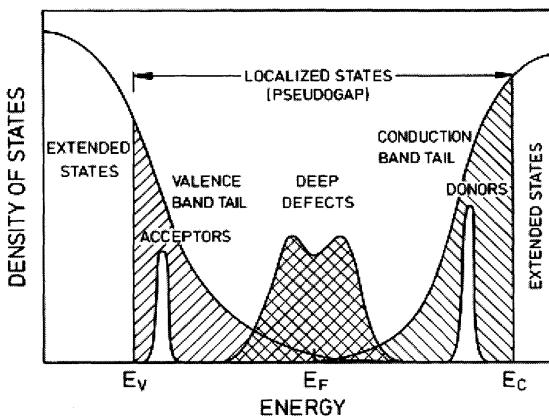


Fig. 7.1. Model of the density-of-states distribution in the energy gap of a-Si:H [11]

A characteristic of the DOS distribution in amorphous semiconductors is tails of localized states extending from both edges deep into the gap. These states are considered to arise from potential fluctuations due to disorder. In addition to these tail states, one expects deep states which originate from specific defects. Such defects may be formed in a random network structure in order to release internal stress but may also arise from unfavorable growth conditions. In a-Si:H the simplest deep defect is an unsaturated bond (Si-db), which has three differently charged states: unoccupied (positive, D^+), singly occupied (neutral, D^0) and doubly occupied (negative, D^-). In undoped a-Si:H the D^0 state is located around midgap, while the D^- state is higher in energy by the correlation energy of about 0.2 eV. Extrinsic states originate from impurities. These films contain hydrogen in a concentration of up to 15 at% and are strongly contaminated with oxygen, nitrogen and carbon in concentrations of the order of 10^{19} cm^{-3} . So far no gap states due to these impurities have been identified. The only known impurity states are those

that arise from substitutional doping by incorporation of elements from group III and V of the periodic table, forming flat donor or acceptor states.

The shape of the absorption edge reveals the general features of the DOS distribution (Fig. 7.2). A comparison of the absorption spectra of amorphous and crystalline silicon reveals a shift of the absorption edge of a-Si:H to higher energy and a strong enhancement of absorption in the visible range of the spectrum, which has often been related to the relaxation of the k -selection rule due to the loss of long-range order. This enhancement of absorption is the basis of important applications such as solar cells, and imaging and optical sensors. An optical gap may be defined in different ways. Rather often, E_{03} and E_{04} are used, which are the photon energies where $\alpha = 10^3 \text{ cm}^{-1}$ and 10^4 cm^{-1} , respectively. Another, more physical way is to use the concept of nondirect optical transitions, which takes into account the loss of the k -selection rule in a disordered structure by plotting the data as $(\alpha h\nu)^{1/2}$ versus $h\nu$ (Tauc plot). By extrapolating the straight line to the energy axis, values for the energy gap of about 1.75 eV are found. A justification for this procedure is that similar values have been obtained from the analysis of transport properties. However, one has to keep in mind that in this case the analysis yields a different quantity, namely the value of the mobility gap $E_C - E_V$. The value of the energy gap appears to be a unique function of the hydrogen concentration C_H in the film [8]. It amounts to 1.2–1.5 eV in evaporated or sputtered material with $C_H = 0$ and increases linearly with C_H in hydrogenated films (1.6–2.0 eV).

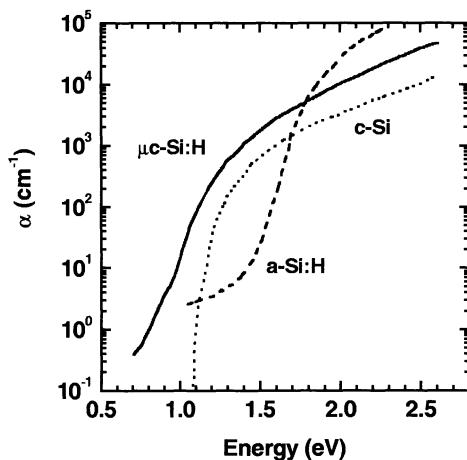


Fig. 7.2. Absorption edges of a-Si:H, monocrystalline Si and μ c-Si:H

Below $\alpha \approx 3 \times 10^3 \text{ cm}^{-1}$, the absorption constant decreases exponentially with energy; $\alpha \sim \exp[-(E/E_0)]$ (Urbach edge). The Urbach parameter E_0 depends on both temperature and the general disorder in the film. The Ur-

bach edge can be related to the exponential bandtails, and it is commonly believed that in a-Si:H the slope of the absorption edge is determined by the slope of the valence bandtail. The values of E_0 (50 meV) for films prepared under optimized conditions agree fairly well with the slope of the exponentially varying density of states of the valence bandtail derived from transport studies. The conduction bandtail is considerably steeper. Time-of-flight experiments have been consistently explained by a slope of about 25 meV [15].

At low photon energies, the absorption curve of a-Si:H levels off and a plateau develops, which is assigned to defect absorption. This defect-related absorption shoulder at low photon energies has quite frequently been used for quantitative defect analysis, applying techniques such as photocurrent spectroscopy and photothermal spectroscopy [16]. The value of α in this energy range is considered a figure of merit for the film quality. In undoped optimized films where, according to electron spin resonance, the concentration of neutral dangling bonds is about 10^{16} cm^{-3} , the value of α at 1.25 eV is in the order of 1 cm^{-1} .

The transport properties are strongly affected by disorder. In a-Si:H, where the density of gap states is low, transport is considered to take place predominantly in extended states above the mobility edges E_C and E_V . The carrier mobility cannot be determined experimentally as in crystalline silicon by the Hall effect, owing to the sign anomaly in amorphous semiconductors. The observation, the explanation of which is still one of the challenges in this research field, is that the sign of the Hall effect is opposite to that expected for the predominant carriers, e.g. positive for electrons. It is estimated that the mobility of the extended states is in the region of $10 \text{ cm}^2/\text{Vs}$. Carriers in localized states can contribute to conduction with much lower mobility by thermally activated tunneling (hopping transport). If the density of states at the Fermi level is high, variable-range hopping may be observed, with a characteristic temperature dependence of the conductivity, $\ln \sigma \propto -(T_0/T)^{1/4}$ [7]. The existence of sharp mobility edges is still controversial and has been questioned as a result of consideration of electron–phonon coupling [11]. Although there is no direct proof for their existence, most experimental results are discussed in terms of this model.

A breakthrough in the physics of amorphous semiconductors was the discovery that a-Si:H can be effectively doped by adding controlled amounts of PH_3 or B_2H_6 to the process gas [4]. In an amorphous semiconductor, the donors deliver their electrons to empty states near the Fermi level. The resultant shift of the Fermi level, therefore, depends on the DOS distribution, and the doping effect in a-Si:H is closely related to the low density of defect states. Figure 7.3 displays typical results of doping. For P doping, the maximum conductivity at 300 K, σ_{RT} , of about $10^{-2} \text{ S cm}^{-1}$, is attained at a concentration of 10^3 – 10^4 ppm PH_3 in the gas phase. At higher doping levels σ_{RT} decreases, presumably owing to the generation of additional defects. For B doping, σ_{RT} decreases at low doping levels, and attains at a high doping level a maxi-

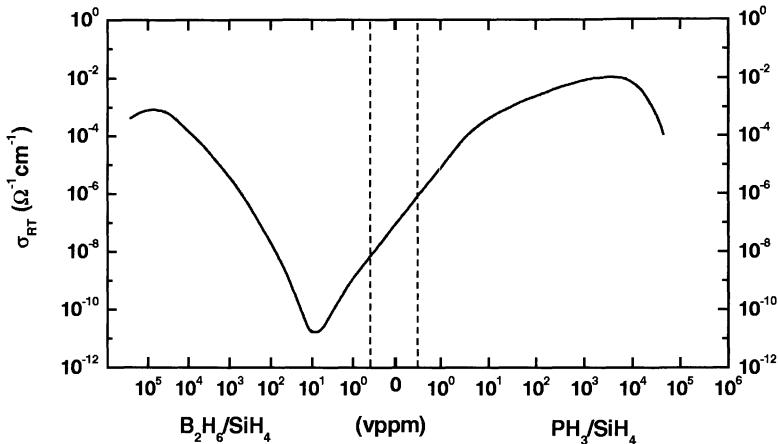


Fig. 7.3. Electrical conductivity at 300 K as a function of the doping level in the gas phase (see e.g. [4, 10, 11])

mum close to $10^{-3} \text{ S cm}^{-1}$. At the minimum, the Fermi level crosses midgap, $E_C - E_F = (1/2)(E_C - E_V)$, and the type of conduction changes from n- to p-type. Using $\sigma_{RT} \approx 10^{-12} \text{ S cm}^{-1}$, the value of the mobility gap can be estimated as 1.7 eV at 300 K. Whereas at high doping levels the results from different laboratories agree fairly well, there are pronounced differences at low doping levels. This is an expression of the influence of the specific deposition conditions on the concentration of deep defects. The details of the doping mechanism have turned out to be very complicated, involving reactions with dangling bonds and hydrogen. Experiments show that the concentration of defects increases strongly with doping, and both the density and the energy position of the defect states are essentially controlled by the position of the Fermi level. This points to a very general defect creation mechanism [16, 17].

7.4 Photoluminescence and Photoconductivity

At temperatures below of 50 K, optimized undoped a-Si:H exhibits photoluminescence (PL) with a quantum efficiency close to unity [18]. The emitted spectrum consists of a single structureless band centered at 1.3–1.4 eV with a width ΔE_{FWHM} of 0.25–0.3 eV (Fig. 7.4). In defect-rich and doped films the emission is quenched and additional structure appears at 0.8–0.9 eV, shown in Fig. 7.4 for the case of boron-doped films. Although the detailed nature of the intrinsic and defect-related emissions is still a matter of debate, it is most widely believed that the intrinsic process occurs by radiative tunneling of electrons and holes localized in the respective bandtails. The photoconductivity (PC) and the photoluminescence show anticorrelated temperature depen-

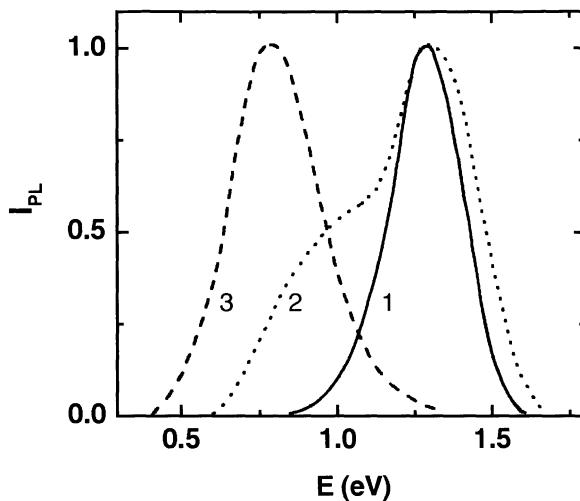


Fig. 7.4. Photoluminescence spectra (normalized intensity I_{PL}) showing the intrinsic emission band at 1.4 eV of undoped a-Si:H (1) and the defect-related band at 0.8 eV of a-Si:H doped with 100 ppm (2) and 1000 ppm (3) of boron

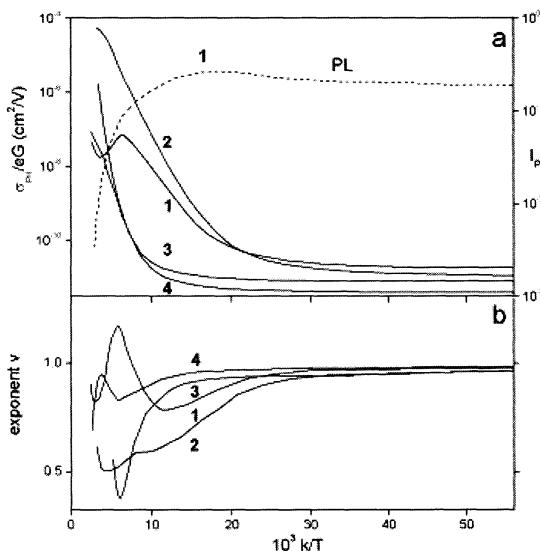


Fig. 7.5. Temperature dependence of **a** the PL intensity I_{PL} and the normalized photoconductivity $\sigma_{PH}/eG = \mu\tau$ at $G = 3 \times 10^{20} \text{ cm}^{-3} \text{ s}^{-1}$ ($\lambda = 525 \text{ nm}$) and of **b** the exponent ν of the intensity dependence $\sigma_{PH} \propto G^\nu$. PECVD a-Si:H: (1) undoped, (2) doped 100 ppm PH₃, (3) doped 10^3 ppm B₂H₆, (4) sputtered a-Si:H (undoped) [19]

dences [19]. Whereas, with rising temperature, the PL intensity is quenched, the photoconductivity increases (Fig. 7.5). This behavior suggests that we should distinguish two temperature ranges of different behavior:

- At $T < 60$ K, the PL quantum efficiency is large (close to unity in good films) and independent of temperature. The photoconductivity σ_{PH} is very small and varies linearly with the light intensity G . Its magnitude normalized by the charge e and generation rate G , $\sigma_{\text{PH}}/eG = \mu\tau$, is about $10^{-11} \text{ cm}^2/\text{V}$ and varies only very little with the quality of the sample. In Fig. 7.5a the product $\mu\tau$ varies only by a factor of 5, whereas the defect densities in the films differ by many orders of magnitude. This transport is quite generally assigned to tunneling of carriers between localized states in the bandtails.
- At $T > 60$ K, the PL intensity is quenched, and σ_{PH}/eG increases strongly. In this temperature range the magnitude of $\sigma_{\text{PH}}(G, T)$ and the kinetics are determined in a complicated way by the defect density, Fermi-level position and other factors. This can be seen from the complicated temperature dependence of the exponent in Fig. 7.5b. Here, transport occurs in extended states above the mobility edges, and Si dangling bonds are considered to be the primary centers for nonradiative recombination [20]. An important feature of phototransport in this regime is that σ_{PH} is tremendously enhanced when the Fermi level is shifted from midgap towards the mobility edges. This effect is particularly pronounced for n-type films when $E_C - E_F$ moves from 0.8 to 0.4 eV. For P-doped samples the mobility–lifetime product $(\mu\tau)_n$ can reach values of up to $10^{-4} \text{ cm}^2/\text{V}$ [19].

7.5 Metastable States

It is a characteristic property of hydrogenated amorphous semiconductors, of which a-Si:H is representative, that deep defects are created by the breaking of weak bonds when band tail states are populated. This can be achieved by exposure to light, carrier injection, strong accumulation in the conduction channel of an a-Si:H thin-film field effect transistor and doping (see above). It has also been reported that defects are generated at elevated temperatures ($T > 420$ K) in thermal equilibrium with the occupancy of tail states [10]. Some of these defects can be frozen in by quenching the film to $T < 420$ K. Such instabilities appear to be an inherent feature of the amorphous structure. The DOS distribution shown in Fig. 7.1 thus is not stable but varies with temperature, doping level and light absorption and during the operation of devices. Most electronic properties therefore depend to some extent on sample history and treatment.

In undoped a-Si:H films of device quality, light exposure causes a reversible decrease of both the dark conductivity and the photoconductivity, indicating a shift of the Fermi energy towards midgap and a decrease of the recombination lifetime [21]. These light-induced changes can be annealed completely at

temperatures of around 170°C. The reason for this Staebler–Wronski effect is a light-induced enhancement of the density of Si dangling bonds to values of typically 10^{17} cm^{-3} . The microscopic mechanism of defect creation is still one of the most important open questions in this research field. Apparently the defect creation is linked not to the absorption process (photodegradation) but to some secondary effect which involves nonequilibrium carriers. This effect could be either recombination or capture of carriers into localized states. It is widely believed now that in device quality material, the effect is intrinsic to hydrogenated films, being the result of disorder and the presence of hydrogen. Most of the results are in accordance with a model where recombination at weak bonds results in bond breaking. The broken bonds are considered to be stabilized by a mechanism which involves the motion of hydrogen atoms. This interpretation is supported in particular by the observation that films deposited by the hot-wire technique are more stable and have a much lower hydrogen content than a-Si:H made by PECVD. Such a participation of hydrogen in the mechanism of defect creation should lead to a spatial correlation between dangling bonds and neighboring SiH bonds, which, however, has not been observed in magnetic resonance studies. Research is being directed towards minimizing the metastability by a proper control of the deposition conditions influencing the microstructure and the concentration of weak bonds (steeper bandtails). There are conflicting observations about deuterated material, a-Si:D, which has been reported to be more stable. It is unclear whether this is due to slower degradation involving SiD bonds or whether this arises from a change in the microstructure of the samples investigated. Some results suggest that the microstructure of the films might indeed be an important factor. Heavy-hydrogen dilution of the process gases has been shown to lead to a more stable material which has microcrystalline inclusions and a low density of microvoids [22]. Recently, similar two-phase materials have been prepared by PECVD using parameters close to the range where powder formation takes place. It was shown that such material has very attractive electronic properties and appears to be more stable [23].

7.6 Amorphous-Silicon Solar Cells

A particularly important field of application was initiated by the first publication about amorphous-silicon solar cells in 1976 [5]. For this application it is most important that the absorption of a-Si:H in the visible spectral region is considerably larger than in c-Si, thus enabling effective absorption of sunlight in a film which is only about 1 μm thick (Fig. 7.2). In the most commonly used superstrate cell, the structure is glass/TCO/p⁺-SiC_x/i-a-Si:H/n⁺-a-Si/metal and the light enters through the glass. The substrate cell uses stainless steel (ss) as the substrate, with an inverted layer sequence ss/Ag-ZnO/a-Si(n⁺)/a-Si(i)/a-Si(p⁺)/metal grid. In both cells the thickness of the active undoped

absorber layer, a-Si:H(i), is about $0.5\text{ }\mu\text{m}$, such that carrier collection occurs predominantly by drift in the electric field. Single-junction cells have been developed with stable efficiencies in the range of 9–10% in the laboratory for small areas. A major problem of the amorphous-silicon technology has been the degradation of amorphous films and devices under illumination. Considerable progress has been made concerning increased stability of solar cells by proper device engineering. The strategy is to use stacks of two or three pin cells. Such an advanced device structure offers two advantages: (1) In a stack, the thickness of the individual cells is reduced, which improves the stability by enhancing carrier collection through an increase of the internal electric field. (2) By using materials with different bandgaps, a tandem structure can be made which leads to a better use of the solar spectrum. It is important that the degradation of the cells appears to saturate as a function of the exposure time (typically after 100 h illumination with AM1), such that stabilized efficiencies can be guaranteed by the manufacturers. Optimized solar cells typically degrade by about 10% from the as-deposited state. In recent years several companies have started commercial production of a-Si:H PV modules with stabilized efficiencies in the range 6–8% on large-area substrates.

Figure 7.6 shows the structure of a triple-junction solar cell from United Solar System Corporation (USSC), the most successful cell of this type so far (see [22]). In this structure, the top cell, which absorbs the high-energy part of the solar spectrum, uses a-Si:H with a bandgap of 1.8 eV as the intrinsic layer. The i-layer of the middle cell is an a-Si/Ge alloy which contains

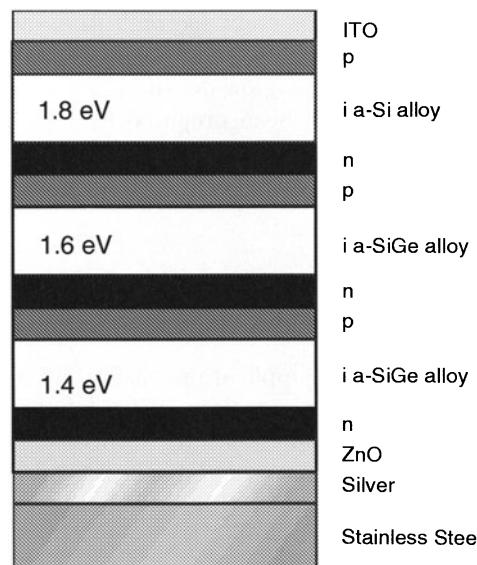


Fig. 7.6. Schematic illustration of a three-junction thin-film solar cell [22]

about 15% Ge, leading to a bandgap of 1.6 eV. The bottom cell, which is designed to absorb the red part of the spectrum, uses an i-layer of an a-Si/Ge alloy with 40–50% Ge and a bandgap of 1.4 eV. A textured Ag/ZnO back-reflector is used to enhance light trapping. Such a structure has led to the highest efficiency (active area) reported so far for laboratory cells ($A = 0.25 \text{ cm}^2$): 15.2% initial, 13% stabilized [22]. The use of high hydrogen dilution in the preparation of the intrinsic films appears to be one of the keys to obtaining these results.

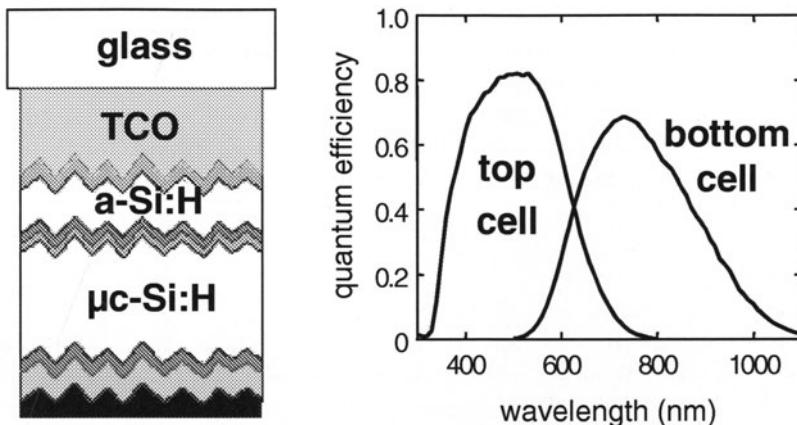


Fig. 7.7. Micromorph solar cell concept: **a** schematic illustration of the cell and **b** quantum efficiencies of the top and bottom cells [26]

The challenges for research at present are the development of materials which can be used in stacked tandem structures with a-Si:H, enhancement of the deposition rate, and improvement of the cell efficiencies and production yields. The a-Si/Ge alloys still have considerably poorer electronic properties than a-Si:H and, in addition, their deposition uses germane as a process gas, which constitutes an important cost factor. Therefore a number of research groups are aiming at replacing the a-Si/Ge alloy with microcrystalline silicon (μ c-Si:H) [24–26]. μ c-Si:H can be prepared by PECVD or HW deposition techniques under conditions which are compatible with those used in the deposition of a-Si:H. The physics of this material is rather complex, owing to pronounced structural heterogeneity. Structural investigations suggest that μ c-Si:H consists of small crystallites with sizes in the order of 5–30 nm which are embedded in columns with a diameter of typically 50–200 nm. A strongly disordered phase exists around the columns, which is often referred to as amorphous silicon. It appears that this structure is largely independent of the method used for preparation. These inhomogeneities determine the physical properties and therefore considerably complicate the interpretation of most

experimental results. A comparison of the absorption edges (Fig. 7.2) suggests that one should use this material for a bottom cell in a tandem concept with a-Si:H. Figure 7.7 shows a schematic illustration of such a cell, made up of a stack of two pin structures with a-Si:H forming the top and μ c-Si forming the bottom cell. The resulting quantum yield reveals the advantage of such a device concept: the top cell, made of a-Si:H, uses the high-energy photons, and the bottom cell, made of μ c-Si:H, uses the low-energy photons. Using this promising approach, stabilized efficiencies of about 10% have been demonstrated.

The industrial activity proves impressively that technologies based on a-Si:H and related alloys have surmounted the barrier to commercialization. The progress achieved in this field is certainly based on a broad understanding of the physics of both materials and devices as a result of 25 years of continuous research effort. The numerous applications of these amorphous semiconductors in photovoltaics, sensors, photodetectors and, in particular, displays drove this research and established the mature technology available today.

References

1. R.C. Chittick, J.H. Alexander, H.F. Sterling: The preparation and properties of amorphous silicon. *J. Electrochem. Soc.* **116**, 77 (1969)
2. W.E. Spear, R.J. Loveland, A. Al-Shabaty: The temperature dependence of photoconductivity in a-Si. *J. Non-Cryst. Solids* **15**, 410 (1974)
3. A.J. Lewis, G.A.N. Connell, W. Paul, J. Pawlik, R. Temkin: Hydrogen incorporation in amorphous germanium. *AIP Conf. Proc.* **20**, 27 (1974)
4. W.E. Spear, P.G. LeComber: Substitutional doping of amorphous silicon. *Solid State Commun.* **17**, 1193 (1975)
5. D.E. Carlson, C.R. Wronski: Amorphous silicon solar cells. *Appl. Phys. Lett.* **28**, 671 (1976)
6. P.G. LeComber, W.E. Spear, A. Gaith: Amorphous silicon field-effect device and possible applications. *Electron. Lett.* **15**, 181 (1979)
7. N.F. Mott, D.E. Davis: *Electronic Processes in Non-Crystalline Materials* (Oxford University Press, Oxford 1979)
8. H. Fritzsche: Characterisation of glow-discharge deposited a-Si:H (review). *Solar Energy Mater.* **3**, 447 (1980)
9. R.K. Willardson, A.C. Beer (eds.): *Hydrogenated Amorphous Silicon, Semiconductors and Semimetals*, vol. 21A-D, ed. by J.E. Pankove (Academic Press, Orlando FL, Cambridge UK 1984)
10. R.A. Street: *Hydrogenated Amorphous Silicon* (Cambridge University Press 1991)
11. H. Overhof, P. Thomas: *Electronic Transport in Hydrogenated Amorphous Semiconductors*, Springer Tracts in Modern Physics 114 (Springer, Berlin, Heidelberg 1989)
12. C. Spinella, S. Lombardo, F. Priolo: Crystal grain nucleation in amorphous silicon. *J. Appl. Phys.* **84**, 5383 (1998)

13. E.P. Donovan, F. Spaepen, D. Turnbull, J.P. Poate, D.C. Jacobson: Calometric studies of crystallization and relaxation of amorphous Si and Ge prepared by ion implantation. *J. Appl. Phys.* **57**, 1795 (1985)
14. O. Nast, S.R. Wenham: Elucidation of the layer exchange mechanism in the formation of polycrystalline silicon by Al-induced crystallisation. *J. Appl. Phys.* **88**, 124 (2000)
15. W.E. Spear: The study of transport and related properties of a-Si:H in transient experiments. *J. Non-Cryst. Solids* **59/60**, 1 (1983)
16. K. Pierz, W. Fuhs, H. Mell: On the mechanism of doping and defect formation in a-Si:H. *Phil. Mag. B* **63**, 123 (1991)
17. R.A. Street, D.K. Biegelsen, J.C. Knights: Doping and the Fermi energy in amorphous silicon. *Phys. Rev. Lett.* **49**, 1187 (1982)
18. R.A. Street: Luminescence and recombination in hydrogenated amorphous silicon. *Adv. Phys.* **30**, 593 (1981)
19. M. Hoheisel, R. Carius, W. Fuhs: Photoconductivity and photoluminescence of a-Si:H at low temperature. *J. Non-Cryst. Solids* **63**, 313 (1984)
20. W. Fuhs, K. Lips: Recombination in a-Si:H films and pin structures studied by electrically detected magnetic resonance. *J. Non-Cryst. Solids* **164/166**, 541 (1993)
21. D.L. Staebler, C.R. Wronski: Optically induced conductivity changes in discharge produced hydrogenated amorphous silicon. *J. Appl. Phys.* **51**, 3262 (1980)
22. J. Yang, A. Banerjee, S. Guha: Correlation of the component cells with high efficiency alloy amorphous silicon alloy triple junction solar cells. *Proc. 2nd World Conference on Photovoltaic Solar Energy Conversion*, Vienna (1998) p. 387
23. P. Roca i Cabarrocas, S. Hamma, S.N. Sharma, G. Viera, E. Bertran, J. Costa: Nanoparticle formation in low-pressure silane plasmas: bridging the gap between a-Si:H and μ c-Si films. *J. Non-Cryst. Solids* **227–230**, 871 (1998)
24. K. Yamamoto, M. Yoshimi, Y. Tawada, S. Fukuda, T. Sawada, T. Meguro, H. Takata, T. Suezaki, K. Koi, K. Hayashi, T. Suzuki, A. Nakajima: Large area thin-film silicon module. *Techn. Digest 12th International Photovoltaic Solar Energy Conference* (2001) p. 547
25. J. Meier, P. Torres, R. Platz, S. Dubail, U. Kroll, J.A. Anna Selvan, N. Pellaton, C. Hof, D. Fischer, H. Keppner, A. Shah, K.-D. Ufert, P. Giannoules, J. Koehler: On the way towards high efficiency thin-film silicon solar cells by the micromorph concept. *Mater. Res. Soc. Symp. Proc.* **420**, 3 (1996)
26. T. Repmann, W. Appenzeller, T. Roschek, B. Rech, O. Kluth, J. Müller, W. Psyk, R. Geyer, P. Lechner: Development a-Si:H/ μ c-Si:H thin-film solar modules using 13.56 MHz PECVD. *Proc. 17th European Photovoltaic Solar Energy Conference* (2001) p. 2836

8 Silicon-on-Insulator and Porous Silicon

J.-P. Colinge

8.1 What is Silicon-on-Insulator?

Silicon is by far the most widely used semiconductor material. It is abundant in the earth's crust and relatively easy to convert into a high-purity single crystal. Unlike some other semiconductor materials, silicon is stable when heated at high temperature, and a well-behaved insulating and passivating material, silicon dioxide, can readily be grown on it. The excellent electrical and chemical properties of thermally grown SiO_2 are probably the most important factor that has made silicon such a successful semiconductor material (concerning devices, see Chap. 18, by Risch).

8.1.1 General Properties of SOI MOS Transistors

Classical silicon devices, such as metal–oxide–semiconductor (MOS) transistors, are made at the surface of silicon wafers that are 700–800 μm thick, but occupy less than the top first micrometer at the surface of the wafer. The remainder of the wafer serves as a mechanical support for the devices and sometimes gives rise to unwanted, parasitic interactions with the devices. In a silicon-on-insulator (SOI) wafer the devices are fabricated in a thin silicon layer. This silicon layer is single-crystal and sits on an insulating material, usually silicon dioxide. Typically the silicon layer thickness ranges from 10 nm to several micrometers, depending on the application, and the silicon dioxide layer thickness ranges between 50 nm and a micrometer. The whole structure rests on a mechanical substrate, typically silicon, although silicon films on glass or quartz substrates are preferred for some applications. The oxide layer between the active top silicon layer and the mechanical silicon substrate is called the buried oxide (BOX). Figure 8.1 shows an MOS transistor made on a silicon, a silicon-on-insulator, and a silicon on glass/quartz substrate. Typical thicknesses used in CMOS applications are indicated. MOS transistors made in an SOI film offer several advantages. In bulk silicon technologies, devices are isolated from one another by reverse-biased PN junctions. As a result there is a capacitance (PN junction capacitance) between the source or the drain and the silicon substrate. The source and drain of SOI devices, on the other hand, are fully isolated from the substrate by a dielectric material. This reduces significantly the capacitance of the source and drain,

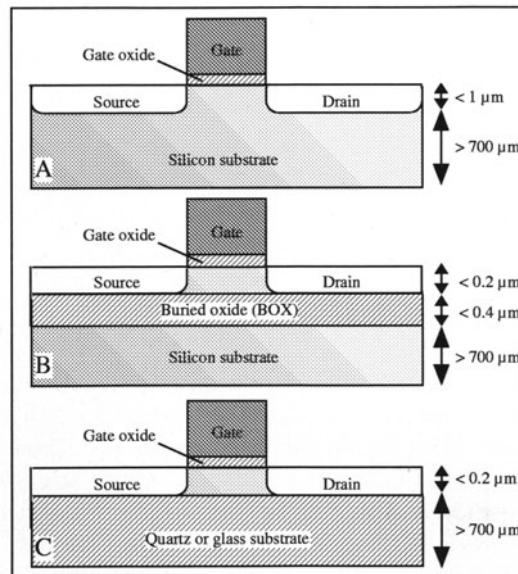


Fig. 8.1. Metal–oxide–semiconductor (MOS) transistor made in A, bulk silicon, B, standard SOI material, and C, silicon on glass/quartz

which allows faster circuit operation. SOI circuits are usually 20–30% faster than their bulk counterparts. Because the transistor body (the part of the device located beneath the gate, between the source and drain) is dielectrically isolated from the silicon substrate, SOI transistors show less body effect than do bulk devices. The body effect reduces current drive and increases the subthreshold swing, which limits the performance of the device under low-voltage operation [1]. As a result, SOI devices have much higher performance in low-voltage, low-power integrated circuit applications than their bulk silicon counterparts have.

Because of the small volume of silicon in the body of an SOI MOSFET, the sensitivity to soft errors caused by alpha particles or cosmic rays is reduced [1, 2]. Furthermore, the reduced area of the source and drain junctions allow device operation at higher temperature than in regular bulk devices [1, 3].

8.1.2 SOI Applications

SOI CMOS technology is now used by many semiconductor manufacturers to fabricate high-speed integrated circuits. Some companies, such as IBM and AMD, fabricate SOI microprocessors using partially depleted (PD) devices, while others, such as Oki, produce BluetoothTM, baseband and RF chips, microcontrollers, DRAMs, SRAMs, and multiplexers for fiber optics using a

Table 8.1. Some VLSI SOI circuits

Company	Circuit	FD/PD	V _{DD}	Performance	Ref.
Samsung	Microprocessor	FD	1.5 V	600 MHz	[5]
IBM	Microprocessor	PD	2 V	580 MHz	[6]
IBM	Microprocessor	PD	1.8 V	550 MHz	[7]
Mitsubishi	DRAM	FD/PD	1 V	46 ns	[8]
Mitsubishi	DRAM	PD	0.9 V		[9, 10]
Samsung	DRAM	PD	< 1.5 V	30 ns	[11]
Hyundai	DRAM	PD	2.2 V	1 Gb	[12]
Samsung	SRAM	FD	0.9 V	20 ns	[13]
Oki	RF and logic	FD			[14]

fully depleted (FD) SOI process [4]. All these products take advantage of the reduced source and drain capacitance of SOI devices, such that a 25–35% speed advantage and a threefold reduction of power consumption, compared with bulk devices, can be achieved. Table 8.1 lists some SOI VLSI circuits.

8.2 SOI Materials

Researchers have spent over 20 years developing reliable techniques to produce SOI wafers. The challenge is to produce a nearly defect-free, device-quality, single crystal of silicon with a diameter of 10 to 30 centimeters, but a thickness of only a fraction of a micrometer. Furthermore, the thin silicon crystal must sit on top of a high-quality amorphous silicon dioxide layer with no mechanical stress or electrically active defects.

8.2.1 Early SOI Materials

During the time period from 1980 to 1990, several techniques were pursued to fabricate SOI substrates. Some of these techniques involved the use of a laser beam, an electron beam, or a focused high-power halogen lamp to melt a thin film of polycrystalline silicon deposited on a silicon dioxide layer. The solidification process of the molten silicon would then be carefully controlled in order to produce a singlecrystal. These techniques received the general name of zone-melting recrystallization (ZMR) techniques. Despite several impressive accomplishments, such as the fabrication of three-dimensional circuits with up to four stacked SOI layers [15], ZMR fabrication of SOI layers was abandoned because of poor yield and reliability.

In another technique called FIPOS (full isolation by porous silicon), the top part of a silicon wafer was transformed into porous silicon using an electrochemical reaction in a hydrofluoric acid bath. Some islands of silicon would be protected from the reaction. The porous silicon was subsequently oxidized and converted into a thermal oxide isolating the silicon islands from the silicon substrate. Although the FIPOS process as such is no longer used, the

formation of a porous silicon layer is used in the ELTRAN™ process to fabricate modern SOI wafers.

A third technique relies on the epitaxial growth of silicon from windows opened in a silicon dioxide layer grown on a silicon substrate. During epitaxy silicon grows laterally on the oxide, thereby forming SOI regions. This technique, called epitaxial lateral overgrowth (ELO), is still used to fabricate three-dimensional device structures [16].

8.2.2 Silicon-on-Sapphire (SOS)

Silicon-on-sapphire (SOS) material was first introduced in 1964 [17]. It is obtained by epitaxial growth of silicon on a (111̄2)-oriented wafer of crystalline alumina (α -Al₂O₃, also called sapphire). The sapphire crystals are produced using Czochralski growth. The sapphire boule is sliced into wafers, which are then subjected to mechanical and chemical polishing. The sapphire wafers receive a final hydrogen etching at 1150°C in an epitaxial reactor, and a silicon film is deposited using pyrolysis of silane at temperatures between 900 and 1000°C. The lattice constants of silicon and (111̄2) sapphire are 0.543 and 0.475 nm, respectively, and the thermal expansion coefficients for silicon and sapphire are 3.8×10^{-6} and $9.2 \times 10^{-6} \text{ K}^{-1}$. Owing to the lattice mismatch between the sapphire and silicon, the defect density in the silicon film is quite high, especially in very thin films. As the film thickness increases, however, the defect density appears to decrease as a simple power law function of the distance from the Si–Sapphire interface. The main defects present in as-grown SOS films are stacking faults and microtwins. Typical defect densities near the Si–Sapphire interface reach values as high as 10^6 planar faults/cm and 10^9 line defects/cm². These account for the low values of the resistivity, mobility, and lifetime near the interface. Because the epitaxial silicon is deposited at high temperature and because the thermal expansion coefficients of silicon and sapphire are different, the silicon film is under compressive stress at room temperature, which reduces the electron mobility but increases the hole mobility.

Several techniques have been developed to reduce both the defect density and the stress in SOS films. The solid-phase epitaxy and regrowth (SPEAR) and the double solid-phase epitaxy (DSPE) techniques are other, more successful, methods for improving the crystal quality of SOS films [18–20]. These techniques employ the following steps. First, silicon implantation is used to amorphize the silicon film, with the exception of a thin superficial layer, where the original defect density is lowest. Then a thermal annealing step is used to induce solid-phase regrowth of the amorphized silicon, the top silicon layer acting as a seed. A second silicon implant is then used to amorphize the top of the silicon layer, which is subsequently recrystallized in a solid-phase regrowth step using the bottom of the film as a seed. In the SPEAR process, an additional epitaxy step is performed after solid-phase regrowth. Using such techniques, substantial improvement of the defect density is obtained. Noise

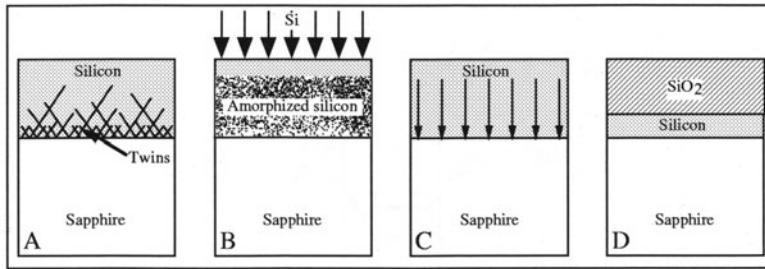


Fig. 8.2. UTSSi process. A, growth of a relatively thick epitaxial silicon film; B, amorphization using silicon ion implantation; C, solid-phase regrowth downward from the defect-free surface; D, thinning of the silicon film by thermal oxidation

in MOS devices is reduced, and the minority carrier lifetime is increased by two to three orders of magnitude [21, 22]. The most recent technique used to produce high-quality SOS is the UTSSi (ultra-thin silicon) process: here, a relatively thick film of silicon is grown on sapphire and, as in the SPEAR process, silicon ion implantation is used to amorphize the film, except in its most superficial part, which is relatively defect-free (Fig. 8.2). Low-temperature annealing is then used to regrow the defect-free silicon downward from the surface through a solid-phase epitaxy mechanism. The silicon film is then thinned to the desired thickness (100 nm) by thermal oxidation and an oxide strip. This process delivers relatively defect-free and stress-free SOS material in which devices with a high effective mobility can be fabricated [22, 23].

8.2.3 SIMOX

The acronym “SIMOX” stands for “Separation by IMplanted OXygen”. The principle of formation of a SIMOX material is very simple (Fig. 8.3), and consists in the formation of a buried layer of SiO₂ by implantation of oxygen ions beneath the surface of a silicon wafer. The buried oxide layer is often referred to as the BOX.

“Standard” SIMOX

The SIMOX technique was invented by Izumi, Doken, and Ariyoshi of NTT in 1978 [24]. In this technique, a high dose of oxygen ions is implanted into a silicon wafer. Ion implantation is traditionally used in the semiconductor industry to introduce doping atoms at the impurity level, and doses higher than a few 10^{15} cm^{-2} are rarely employed. In the SIMOX technique implanted oxygen atoms are used to synthesize a new material, namely silicon dioxide. As a result, a very high dose of oxygen ions (typically $1.8 \times 10^{18} \text{ cm}^{-2}$ at 200 keV in the “standard” SIMOX process) has to be implanted to form the

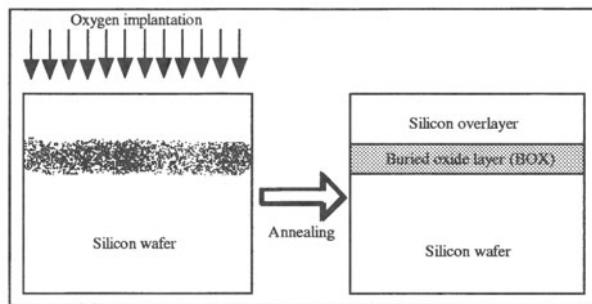


Fig. 8.3. The principle of SIMOX: a heavy-dose oxygen implantation into silicon followed by an annealing step produces a buried layer of silicon dioxide below a thin, single-crystal silicon overlayer

buried oxide layer. Stoichiometric SiO_2 contains 4.4×10^{22} oxygen atoms/cm³. Therefore, the implantation of 4.4×10^{17} atoms/cm² should be sufficient to produce a 100 nm thick buried oxide layer. Unfortunately, owing to the statistical nature of ion implantation, the oxygen profile in silicon does not have a box shape, but instead a skewed Gaussian profile, and the implanted atoms are spread over more than 100 nm, such that the SiO_2 stoichiometry is not reached (Fig. 8.4). If the wafer is annealed after implanting an oxygen dose that is too low, oxide precipitates form at a depth equal to the depth of maximum oxygen concentration, but no continuous layer of SiO_2 is produced. Experiments show that a dose of 1.4×10^{18} cm⁻² must be implanted (at an energy of 200 keV) in order to create a continuous buried oxide layer. The standard dose which is most commonly used is 1.8×10^{18} cm⁻², which produces a 400 nm thick buried oxide layer upon annealing. Figure 8.4 describes the evolution of the profile of oxygen atoms implanted into silicon with an energy of 200 keV. At low doses, a Gaussian oxygen profile is obtained. When the dose reaches 1.4×10^{18} cm⁻², stoichiometric SiO_2 is formed (66 at.% of oxygen and 33 at.% of silicon), and further implantation does not increase the peak oxygen concentration, but instead broadens the overall profile (i.e. the buried oxide layer becomes thicker). This is possible because the diffusivity of oxygen in SiO_2 is high enough for the oxygen to readily diffuse to the Si– SiO_2 interface, where oxidation occurs. The dose at which the buried oxide starts to form ($\cong 1.4 \times 10^{18}$ cm⁻²) is called the “critical dose”.

The temperature at which the implantation is performed is also an important parameter which influences the quality of the silicon overlayer. The oxygen implantation step amorphizes the silicon which is located above the projected range. If the temperature of the silicon wafer during implantation is too low, the silicon overlayer becomes completely amorphized, and it forms polycrystalline silicon upon further annealing, an undesirable ef-

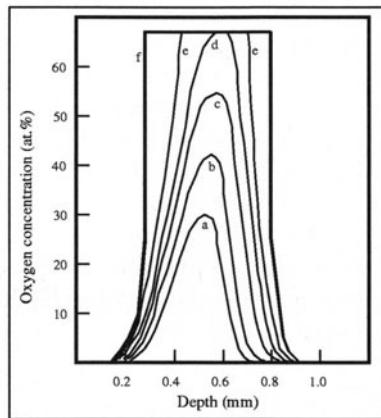


Fig. 8.4. Evolution of the oxygen concentration profile with the implanted dose for an implantation energy of 200 keV: (a) $4 \times 10^{17} \text{ cm}^{-2}$, (b) $6 \times 10^{17} \text{ cm}^{-2}$, (c) 10^{13} cm^{-2} , (d) $1.2 \times 10^{13} \text{ cm}^{-2}$, (e) $1.8 \times 10^{13} \text{ cm}^{-2}$, and (f) $2.4 \times 10^{13} \text{ cm}^{-2}$

fect. When the implantation is carried out at higher temperatures (above 500°C), the amorphization damage anneals out during the implantation process (“self-annealing”), and the single-crystal nature of the top silicon layer is maintained. The silicon overlayer, however, is highly defective and the ion implantation step must be followed by a high-temperature anneal step to improve the quality of both the BOX and the silicon layer.

Since every implanted oxygen atom must traverse the top silicon layer (the future silicon-on-insulator layer), a large number of defects is created. The typical defect density of early SIMOX layers was in excess of 10^9 defects

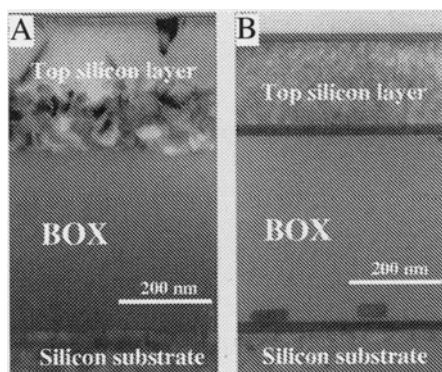


Fig. 8.5. A, SIMOX produced in 1985, 10^9 defects/ cm^2 ; B, SIMOX produced in 1998, less than 1000 defects/ cm^2 [27]

per square centimeter. Constant improvement to SIMOX was achieved by many research groups worldwide. In a nutshell, these improvements consist of maintaining the wafer at a temperature where most defects self-anneal during implantation, and performing a subsequent thermal treatment at high temperature (1350°C) in an appropriate ambient (argon + 2% oxygen) to allow the stabilization and densification of the BOX, as well as the removal of oxide precipitates and other defects in the top silicon layer. Figure 8.5 shows TEM cross sections of SIMOX samples fabricated in 1985 and 1998. The improvement in quality of the top silicon layer can readily be appreciated. The small silicon inclusions at the bottom of the BOX are characteristic of the standard SIMOX material [25, 26].

Low-Dose SIMOX

In 1990 Nakashima and Izumi proposed to reduce the implanted oxygen dose to drastically reduce the dislocation density in the silicon overlayer film. They found that the dislocation density drops drastically as the dose is reduced below $1.4 \times 10^{18} \text{ cm}^{-2}$ at an implantation energy of 180 keV (Fig. 8.6) [28].

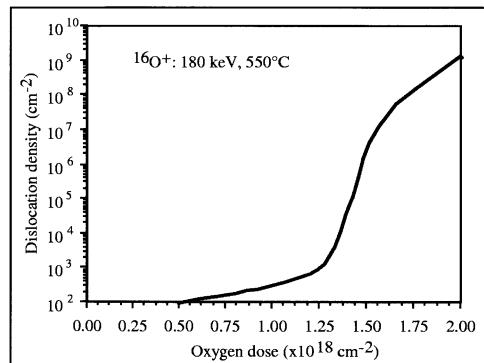


Fig. 8.6. Evolution of dislocation density in the silicon overlayer with implanted oxygen dose [29]

Besides the reduction of the defect density in the silicon layer, there are other motivations for reducing the oxygen dose used to produce SIMOX material. Firstly, the total-dose radiation hardness of thin buried oxides is expected to be better than that of thicker ones. Secondly, direct fabrication of thin buried oxides is attractive for thin-film device applications. Finally, the production cost of a SIMOX wafer is proportional to the implanted dose. A potential additional benefit from this technique is the reduction of contamination of the wafers, since the introduction of impurities (carbon

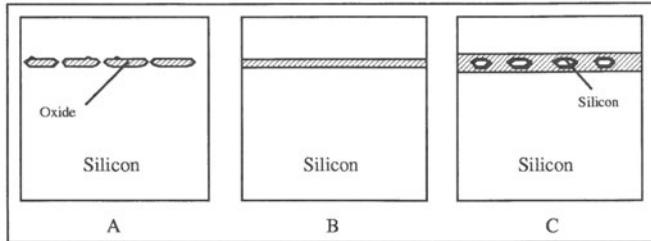


Fig. 8.7. Evolution of the buried oxide structure for a dose of (A) 3×10^{17} , (B) 4×10^{17} , (C) 5×10^{17} $\text{O}^+ \text{ cm}^{-2}$ and an energy of 120 keV [34]

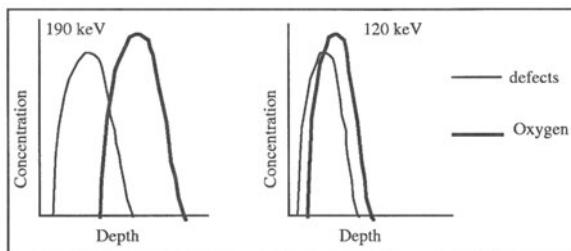


Fig. 8.8. Defect and oxygen ion concentrations produced by 190 and 120 keV implants

and heavy metals) is proportional to the implanted oxygen dose. Low-dose SIMOX is obtained by implanting O^+ ions within a narrow dose window around 4×10^{17} atoms/ cm^{-2} [30]. With a single implantation and a 6 hour anneal at 1320°C , a continuous BOX having a thickness of 80 nm is formed.

Figure 8.7 presents the structure of the buried oxide versus dose around the process window for an implant energy of 120 keV [31]. At a dose of 3×10^{17} cm^{-2} , isolated oxide precipitates are formed. For a dose of 5×10^{17} cm^{-2} , silicon precipitates form in the BOX. Only doses within a very narrow process window around 4×10^{17} cm^{-2} produce a continuous, precipitate-free BOX. The choice of the implantation energy is a critical parameter. At 190 keV, which is close to the standard SIMOX energy, some SiO_2 islands are found in the silicon overlayer. When the peak of implant defect generation and the projected range of the oxygen ions are distinct, two precipitation sites can occur, and oxide precipitates can form at both the oxygen projected range and the peak of defect generation. A reduction of the implant energy to 120 keV is sufficient to merge the two precipitation sites and obtain a single and continuous BOX (Fig. 8.8). The use of a lower implantation dose significantly reduces the defect density; dislocation densities on the order of 300 cm^{-2} are found in low-dose SIMOX material. From an economic point of view, the use of low-dose implantation is obviously advantageous, since the throughput

of the oxygen implanter is inversely proportional to the implanted dose. The lowest-dose SIMOX material reported uses a $2 \times 10^{17} \text{ O}^+/\text{cm}^2$ implant, which produces a 56 nm thick BOX [32]. An empirical relationship between implant dose and energy for the production of thin buried oxide layers has been experimentally established, and it has been shown that for doses ranging between 2 and $6 \times 10^{17} \text{ O}^+ \text{ cm}^{-2}$, the use of an implant energy equal to E (keV) $\cong 30 \times D$, where D is expressed in 10^{17} cm^{-2}), gives the best results [33].

ITOX

It is possible to increase the thickness of the BOX produced by low-dose oxygen implantation; high-temperature (1350°C) oxidation of a low-dose SIMOX wafer causes an increase of the thickness of the buried oxide (Fig. 8.9).

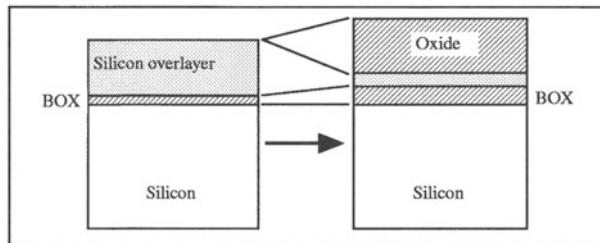


Fig. 8.9. Principle of internal thermal oxidation (ITOX) [35–37]

This phenomenon is called high-temperature internal oxidation (ITOX). As long as the thermal oxide grown on the silicon overlayer is thinner than 500 nm, there exists a linear relationship between the thickness of this oxide layer (t_{ox}) and the thickness increase of the buried oxide (Δt_{BOX}) : $\Delta t_{\text{BOX}} = 0.06 t_{\text{ox}}$. The internal oxide grows at the expense of the bottom of the silicon overlayer. High-temperature internal oxidation has been shown to significantly improve the roughness of the interface between the silicon overlayer and the BOX and to densify the oxide itself.

MLD SIMOX

In low-dose SIMOX, the formation of a continuous buried oxide layer is possible if the peak oxygen concentration is located in highly defective silicon. This observation is the basis for the modified low-dose (MLD) SIMOX process [38].

The MLD process overcomes the problem of oxide continuity encountered during low-dose SIMOX processing. To promote the formation of an ultrathin buried oxide during post-implantation annealing, the implantation process is modified to produce a microstructure which promotes coalescence of the oxygen into a continuous layer. This is accomplished by performing a two-step

implant. Firstly, oxygen is implanted at a dose of $3 \times 10^{17} \text{ cm}^{-2}$ and an energy of 150 keV at a temperature of 525°C. These dose and temperature values ensure minimal generation of defects in the silicon overlayer. Unfortunately, the defects generated by the implant are not located at the peak of the oxygen concentration. To introduce additional defects, a final increment of the dose is implanted near room temperature. This dose (10^{15} cm^{-2}) is chosen to selectively amorphize the region near the depth of the oxygen peak concentration, which yields a highly defective layer during subsequent annealing. This layer provides a template or guide upon which the oxide forms. Buried oxides prepared in this way have been shown to be continuous and without silicon inclusions [39]. It is worth noting that an additional implant of oxygen after the formation of the oxide can be used to improve the stoichiometry of the oxide and improve its electrical properties. It has been shown that the implantation of oxygen doses ranging between 10^{15} and 10^{17} cm^{-2} into the BOX of an already formed SIMOX structure (either “standard” or low dose) increases the density of the oxide and reduces the trapping of charge in the oxide when the material is exposed to ionizing radiations [40]. The thickness uniformity of MLD SIMOX wafers is better than 2–3 nm (6σ) and the RMS surface roughness of the SOI is 0.12 nm ($1 \times 1 \mu\text{m}^2$ AFM scan). The throughput of a high-current oxygen implanter (beam current = 80 mA) is approximatively 30 000 300 mm MLD SIMOX wafers/year [41].

Related Techniques

It is possible to form a buried oxide layer without actually implanting oxygen. As we have seen earlier, the simultaneous presence of defects and oxygen in silicon can result in the formation of a continuous buried oxide layer. In 2001, Ogura, [42], published a description of the formation of a continuous BOX obtained by implantation of light ions (H^+ or He^+) and subsequent annealing in an oxygen-containing ambient. The implantation of hydrogen or helium ions creates a defective layer near the projected range of the ions. Few defects, however, are created in the rest of the silicon, including in the future silicon overlayer, since H^+ and He^+ are light ions. The reported implant conditions are H^+ , $5 \times 10^{16} \text{ cm}^{-2}$, 45 keV, and He^+ , $1-5 \times 10^{17} \text{ cm}^{-2}$, 45 keV, all implanted at room temperature. Upon annealing in an argon/oxygen ambient at temperatures ranging from 1200 to 1350°C, oxygen diffuses through the top of the silicon layer and an internal oxidation process takes place which forms a buried oxide layer where the defects were present (Fig. 8.10). This technique makes it possible to produce a SIMOX-like SOI structure without the need for oxygen implantation and with less damage to the silicon overlayer. It is worth noting that oxygen implantation can be used to create the defective layer, as an alternative to hydrogen or helium [42].

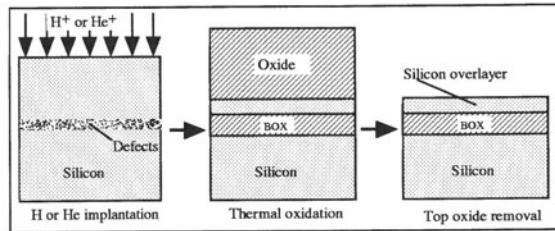


Fig. 8.10. Buried-oxide formation by light-ion implantation and annealing in oxygen atmosphere [43]

8.2.4 Wafer Bonding and Etch-Back (BESOI)

The expression “wafer bonding” refers to the phenomenon that mirror-polished, flat, clean wafers of almost any material, when brought into contact, are locally attracted to each other by van der Waals forces and adhere or “bond” to each other.

The bonding at room temperature is usually relatively weak. Therefore, for many applications, room-temperature-bonded wafers must undergo a heat treatment to strengthen the bonds across the interface. After wafer bonding, one of the wafers is subsequently polished or etched down to a thickness suitable for SOI applications. The other wafer serves as a mechanical substrate, and is called the handle wafer (Fig. 8.11)

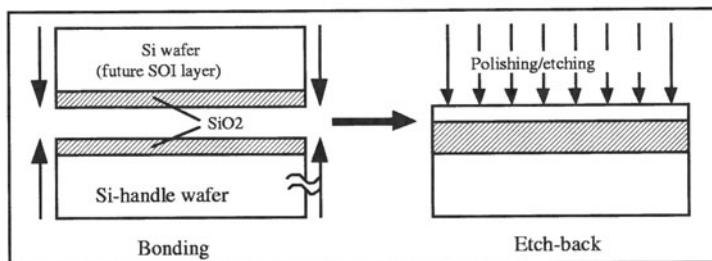


Fig. 8.11. Bonding of two oxidized silicon wafers (*left*), and polishing/etching back of one of the wafers

Hydrophilic Wafer Bonding

When two flat, hydrophilic surfaces such as oxidized silicon wafers are placed against one another, bonding occurs naturally, even at room temperature. The contacting force is caused by the attraction of hydroxyl groups (OH^-)

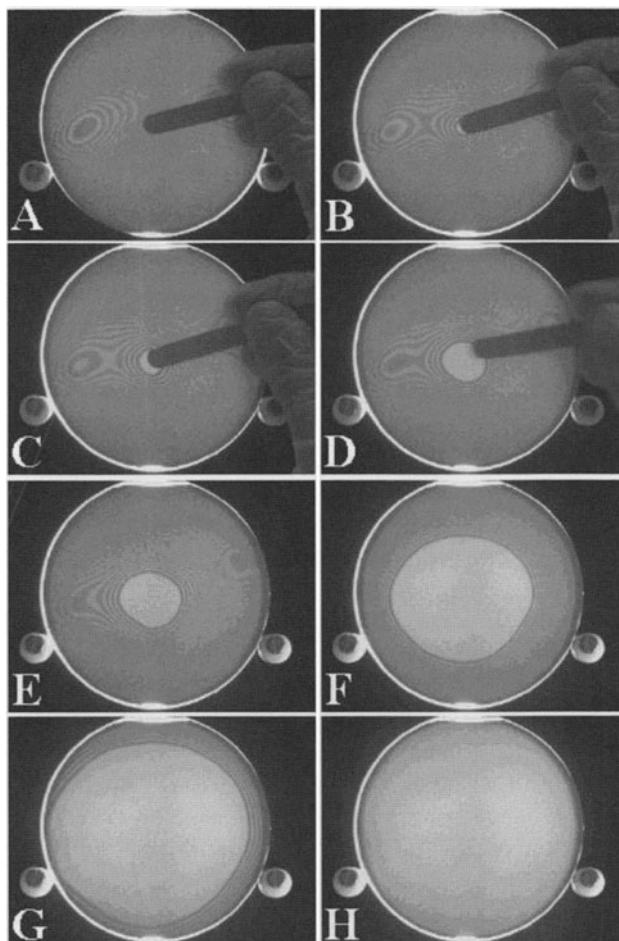


Fig. 8.12. Infrared transmission imaging, showing the propagation of a bonding wave between two oxidized 100 mm silicon wafers (from A to H). The time between successive pictures is approximately 0.5 second. (Courtesy K. Hobart)

adsorbed on the two surfaces. This attraction propagates from the first site of contact across the whole wafer in the form of a “contacting wave” with a speed of several cm/s. Figure 8.12 shows the propagation of a bonding wave between two oxidized 100 mm silicon wafers.

Directly after room temperature bonding, the adhesion between the two wafers is determined by van der Waals interactions or hydrogen bridge bonds and is one or two orders of magnitude lower than what is typical for covalent bonding. For most practical applications, a higher bond energy is required, which may be accomplished by an appropriate heating step, which frequently for commercial SOI production is performed at temperatures as

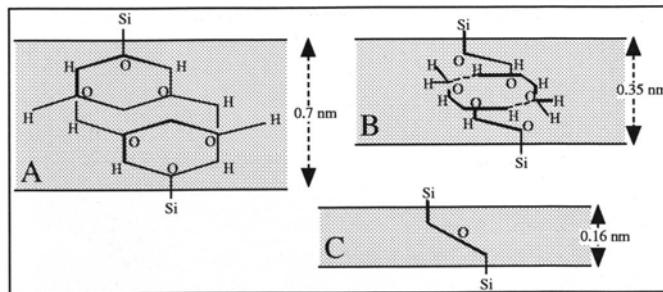


Fig. 8.13. Stengl's proposed model for silicon wafer bonding at different temperatures. A, room temperature, $\text{SiOH}:(\text{OH}_2)_2:(\text{OH}_2)_2:\text{HOSi}$; B, $T = 200^\circ\text{C}$, $\text{SiOH}:\text{HOSi} + (\text{H}_2\text{O})_4$; C, $T > 700^\circ\text{C}$, $\text{SiOSi} + \text{H}_2\text{O}$ [48]

high as 1100°C. Stengl et al. developed a mechanistic model for direct wafer bonding for oxidized wafers which describes the bonding chemistry at different temperatures [44]. When silica surfaces are hydrated, water molecules cluster on the oxidized wafer surface as shown in Fig. 8.13. When two such surfaces are brought into contact, hydrogen bonding occurs via the adsorbed water. At temperatures above 200°C, the adsorbed water separates from the SiOH group and forms a tetramer water cluster. At temperatures greater than 700°C, the water clusters decompose and diffuse away, leaving Si–O–Si bonds. Maszara's experimental findings agree with Stengl's model, where reaction bonding proceeds by two different reactions [45]. Maszara confirmed the mechanism by measuring surface energies of bonded wafers using the crack propagation method. The bond strength for room-temperature contact-bonded wafers varies between 60 and 85 mJ/m², which is consistent with the surface energy of silica bonded through hydrogen bonding. In addition, the surface energy increases at a transition temperature of 300°C. This is the temperature where hydrogen bonds begin to convert to Si–O–Si bonds. The surface energy is constant for bonded wafers annealed in the region from 600°C to 1100°C for anneal times between 10 seconds and 6 hours, and reaches values in excess of 1000 mJ/m². Maszara, [45], concludes that the bonding process does not involve mass transport in that temperature range. Rather, the bond strength is limited in that regime by the amount of contacted area of the bonded wafers, which is a function of how well the wafers can deform elastically at a specific temperature. The kinetics of the deformation are so fast that the bond strength appears to be a function of temperature only. For temperatures greater than 1100°C, the bond energy does increase with time of anneal, but this is due to the viscous flow of the oxide at these high temperatures [46]. It is worth noting that the bond strength of room-temperature-bonded wafers can be significantly increased by plasma treating (in an oxygen plasma) the wafers prior to bonding [47].

Etch-Back

After bonding of the wafers has been carried out, the top wafer has to be thinned down from a thickness, say, of 600 µm to a few micrometers or less in order to be useful for SOI device applications. Two basic thinning approaches can be used: grinding followed by chemico-mechanical polishing, and grinding followed by selective etch-back. The grinding operation is a rather crude but rapid step, which is used to remove all but the last several micrometers of the (top) bonded wafer. The thinning method using chemico-mechanical polishing is cheap, but its use is, so far, limited to the fabrication of rather thick SOI films because of the absence of an etch stop. Much more accurate are the techniques using, after initial grinding, a chemical etch-back procedure with etch stop(s). The etch stop is usually obtained by creating a doping concentration gradient at the surface (i.e. right next to the oxide layer used for bonding) of the top wafer. For instance, in the double-etch-stop technique, a lightly doped wafer is used, and a P⁺⁺ layer is created at its surface by ion implantation. Then, a lightly doped epitaxial layer is grown onto it. This epitaxial layer will be the SOI layer at the end of the process. After grinding, two chemical etch steps are used. First, a potassium hydroxide solution [49] is used to etch the substrate; this stops in the P⁺⁺ layer. Then a 1:3:8 HF:HNO₃:CH₃COOH etch is used to remove the P⁺⁺ layer. The combined selectivity of the etches is better than 10 000:1. The final thickness uniformity of the SOI layer depends on the uniformity of the thickness of the silicon grown epitaxially, as well as on the uniformity of the P⁺⁺ layer formation, but thickness standard deviations better than 12 nm can be obtained.

8.2.5 Smart-Cut®

The Smart-Cut® process combines ion implantation technology and wafer bonding to transfer a thin surface layer from a wafer onto another wafer or an insulating substrate [50, 51]. It consists in a succession of three steps: implantation of gas ions (usually hydrogen), bonding to a stiffener, and a thermal annealing (Fig. 8.14):

- Ion implantation of hydrogen ions into an oxidized silicon wafer (called the “seed wafer”). The implanted dose is on the order of $5 \times 10^{16} \text{ cm}^{-2}$. At this stage, microcavities and microbubbles are formed at a depth equal to the implantation range (R_p). The wafer is preferably capped with thermally grown SiO₂ prior to implantation. This oxide layer will become the buried oxide of the SOI structure at the end of the process.
- Hydrophilic bonding of the seed wafer to another oxidized silicon wafer, called the “handle wafer”, is performed.
- A two-phase heat treatment of the bonded wafers is then carried out. During the first phase, which takes places at a temperature around 500°C, a crystalline rearrangement and coalescence of the microbubbles into larger

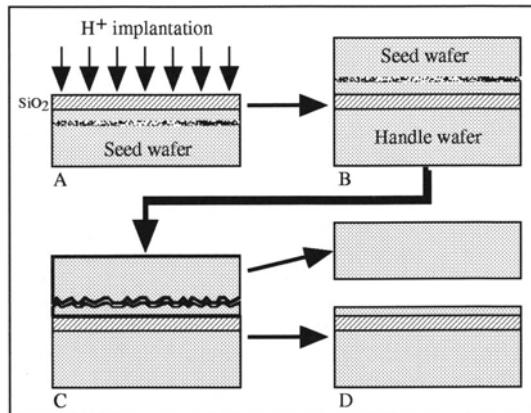


Fig. 8.14. The Smart-Cut® process. A, hydrogen implantation; B, wafer bonding; C, splitting of wafer A; D, polishing of both wafers. Wafer A is recycled as a future handle wafer

structures occurs in the hydrogen-implanted region of the seed wafer. The hydrogen pressure builds up in the growing cavities and eventually the seed wafer splits into two parts: a thin layer of monocrystalline silicon, which remains bonded to the handle wafer, and the remainder of the seed wafer, which can be recycled for later. The basic mechanism of the wafer splitting upon hydrogen implantation and thermal treatment is similar to the surface flaking and blistering of materials exposed to helium or proton bombardment. During the anneal, the average size of the micro-cavities increases. This size increase takes place along a (100) direction (i.e. parallel to the wafer surface) and an interaction between cavities is observed, which eventually results in the propagation of a crack across the whole wafer. This crack is quite parallel to the bonding wafer. The second heat treatment takes place at a higher temperature (1100°C) and is aimed at strengthening the bond between the handle wafer and the SOI film.

Finally, chemo-mechanical polishing is performed on the SOI film to give it the desired mirror-like surface. This layer exhibits significant micro-roughness after wafer splitting, such that a final touch-polish step is necessary. This polishing step reduces the surface roughness to less than 0.15 nm and consumes a few hundred angstroms of the SOI film. Note that because the seed wafer can be recycled, only $N + 1$ silicon wafers are required to produce N SOI substrates.

The basic mechanisms involved in the Smart-Cut® process are described in the following sections.

Hydrogen/Rare Gas Implantation

Implantation of rare gas ions into materials has long been known to lead to the formation of blisters at the material surface. The implantation of alpha particles (He^{++} ions) produced by nuclear reactions into the vacuum walls of fusion reactors causes the wall surface to flake and become covered with blisters [52]. Implantation of a variety of gases into different materials has been shown to cause blister formation, e.g. Ar^+ in Ge [53], Ar^+ in Si [54], H^+ in GaP [55], He^+ in Mo and Nb [56], He^+ in Ni [57], He^+ and Ne^+ in Al [58], Ar^+ and Xe^+ in Si [59], and H^+ in Si [60]. In early blistering experiments, high-fluence implantation (10^{17} – 10^{18} cm^{-2}) was used to create gas-filled cavities at a depth near the projected range of the implanted species. If the dose is high enough the gas pressure in the cavities leads to mechanical deformation of the material above the cavity and forms a blister. If the pressure in the cavity is sufficiently large, the lids of the blisters can break and flaking is observed at the material surface.

It is also possible to form blisters by implanting a moderate dose ($< 10^{17} \text{ cm}^{-2}$) of gas ions and then performing a thermal annealing step to promote the coalescence of small gas-containing defects into larger ones. For the fabrication of SOI material, it is suitable to use hydrogen implantation rather than He or any other rare gas because the low mass of hydrogen ions results in the deposition of little energy in the region situated above the projected range of the ions. As a result, few defects are created in the silicon surface layer (the future SOI film) and most defects are situated at a depth near the projected range [61].

In the particular case of hydrogen (proton) implantation into silicon, the defects created by implantation, or hydrogen-related cavities (HRCs), consist of a mix of hydrogen-decorated vacancies, vacancy clusters, and platelets [62]. Platelets are flat, disk-shaped microcavities containing hydrogen. Their thickness is approximately 2 lattice parameters (1 nm), their diameter is approx-

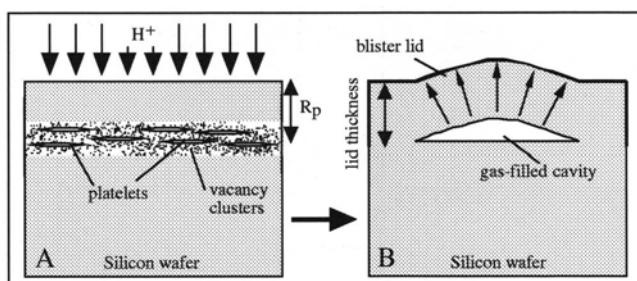


Fig. 8.15. A: hydrogen implantation and formation of defects (vacancies, vacancy clusters, and platelets). B: formation of a blister upon annealing. R_p is the projected range of the implanted ions

imately 10 nm, and they are mainly oriented along (100) planes parallel to the (100) surface of the wafer. (Fig. 8.15). Typical implantation doses range between 1×10^{16} and 7×10^{16} H^+/cm^2 . Upon annealing, typically above 500°C, the hydrogen atoms in the HRCs sever their bonds with silicon atoms and diffuse in the silicon. The hydrogen atoms aggregate in the larger defects, forcing them to grow in size, through a mechanism called Ostwald ripening: the hydrogen atoms lost by the small cavities are captured by larger ones, such that larger cavities grow at the expense of the smaller ones [63, 64]. Molecular hydrogen migrating into the larger cavities causes a buildup of pressure and the formation of blisters [65, 66]. The coalescence of small defects into blisters in silicon implanted with hydrogen has been shown to occur at temperatures as low as 250°C (H_2^+ , 160 keV, $5 \times 10^{16} \text{ cm}^{-2}$) [67].

Bonding to a Stiffener

The Smart-Cut® process is based on harnessing the destructive forces produced during blister formation to produce a thin silicon film. This is accomplished by attaching a stiffener to the implanted silicon wafer. Usually the stiffener is an oxidized silicon wafer, called the “handle wafer” (Fig. 8.16A), but glass, quartz, and other materials can be used as handle substrates as well. The implanted wafer is usually called the “seed wafer” because it is the wafer from which the thin SOI film will come. An annealing step is then performed to both strengthen the bond between the two wafers and to make it possible for the smaller HRCs to coalesce into larger, hydrogen-filled structures called “cracks” or “microcracks”.

These cracks correspond to the gas-filled cavities that form blisters in the absence of a stiffener. The role of the stiffener is to prevent mechanical deformation of the thin silicon layer between the defect region and the SiO_2 layer. In addition, the presence of the stiffener provides a restoring force

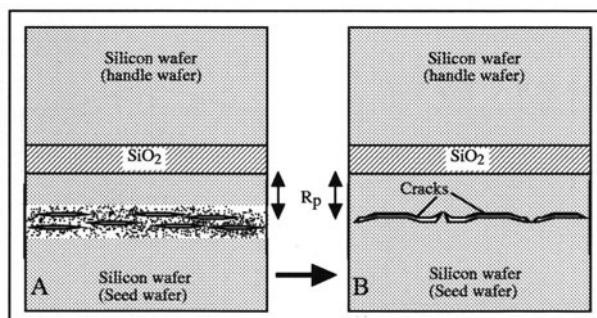


Fig. 8.16. Formation of cracks near the projected range of a silicon wafer implanted with hydrogen. A, bonding of the handle wafer (stiffener) to the seed wafer; B, formation of a crack network near the projected range upon annealing

that opposes the vertical liftoff that leads to blistering and drives lateral crack propagation instead. As a result, a network of horizontal cracks filled with pressurized hydrogen is created, which creates a “perforated-line”-like separation between the top part of the seed wafer and the rest of that wafer.

Annealing

The annealing step is responsible for removing hydrogen atoms from the vacancy complexes and the platelets and feeding them to the growing microcracks. Many different experimental techniques have been used to investigate the coalescence of small HRCs into larger bubbles, blisters, or microcracks. Time-to-blister and time-to-splitting experiments have been carried out in order to extract the activation energies involved in the Smart-Cut process. The time to splitting of the seed wafer can be accurately measured. It depends of course on the annealing temperature, but also on the implantation conditions (dose and energy, doping species and concentration in the silicon, and bonding parameters). Two types of activation energies can be seen from an Arrhenius-type plot of $1/(t \text{ in s})$ versus $1/kT$ (Fig. 8.17). At high temperatures the activation energy E_a is 0.5 eV , which is very close to the activation energy for the diffusion of hydrogen in silicon (0.48 eV). At lower temperature the activation energy is approximately 2.2 eV . This corresponds to the sum of activation energies for the diffusion of hydrogen (0.48 eV) and for breaking the H–Si bonds in the HRCs ($1.8 \pm 0.2 \text{ eV}$). Thus, at high temperatures the coalescence of the bubbles is diffusion-limited, while at lower temperatures ($< 500^\circ\text{C}$) the reaction is limited by both the extraction of the hydrogen from the hydrogen-related cavities and hydrogen diffusion.

The Ostwald ripening mechanism by which hydrogen migrates from the smaller defects into the larger ones has been confirmed by combining infrared spectroscopy and forward recoil scattering (FRS) on wafers annealed

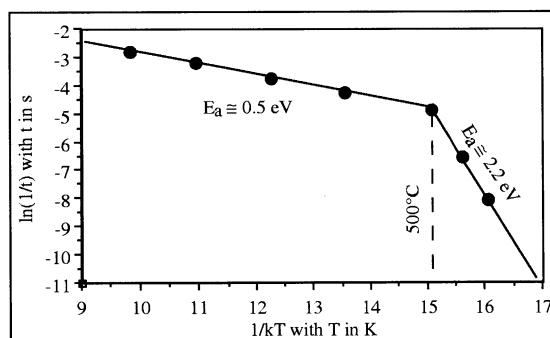


Fig. 8.17. Arrhenius plot of the time to splitting versus annealing temperature. The implant conditions were H^+ , 69 keV , $6 \times 10^{16} \text{ cm}^{-2}$ [68]

without a stiffener. Infrared spectroscopy detects Si–H bonds in the HRCs, while FRS measures the total hydrogen concentration (hydrogen in the bonds and molecular hydrogen in the forming blisters). FRS shows that there is no outdiffusion of hydrogen from the silicon as long as the annealing temperature stays below 500°C, and thus that the total amount of hydrogen remains constant (Fig. 8.18). Infrared spectroscopy indicates that the amount of hydrogen in the Si–H bonds decreases with annealing temperature between 200 and 500°C. The hydrogen in the Si–H bonds disappears almost totally for annealing temperatures above 500°C. This clearly demonstrates the conversion of the hydrogen trapped in the HRCs into an unbound form, i.e. molecular hydrogen. This hydrogen increases both the size of and the pressure in the defects where it accumulates.

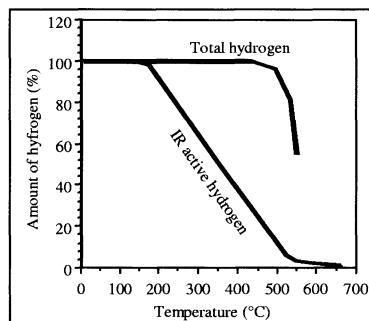


Fig. 8.18. Evolution of hydrogen in Si–H bonds and total amount of hydrogen versus annealing temperature. The annealing time was 30 min [69]

Transmission electron microscopy has been used to measure the size and the density of the microcavities with annealing time ($T = 450^\circ\text{C}$). Such a study shows that the size of the cavities increases with annealing time, while their density decreases. More importantly, the total volume of the cavities (the product of their volume with their density) remains constant with annealing time, once again indicating that the total amount of hydrogen is constant and that the gas moves from the smaller defects into the larger ones (Ostwald ripening) [70].

Splitting

Splitting of the seed wafer takes place naturally at the end of the annealing process, when the silicon is sufficiently weakened near the projected-range depth by the network of microcracks and the buildup of pressure in them. The duration of the splitting itself is extremely short (probably less than

one millisecond) and generates a small audible noise. It is, however, possible to induce splitting of the seed wafer by other means, once the silicon has been weakened sufficiently by hydrogen implantation and some annealing. For instance, splitting can be obtained by dipping the wafer pair into liquid nitrogen [71] or applying a mechanical force to the edge of the seed wafer [72, 73]. Splitting can also be obtained by supplying microwave energy to the wafer (2-minute microwave “anneal” at 900 W, 2.45 GHz) after bonding the implanted wafer and annealing the bonded pair at 150°C [74]. After splitting, the SOI wafer receives a final “touch” mechanical-chemical polish step to reduce surface roughness and achieve the desired surface flatness for the fabrication of SOI devices. Exposure to an HCl + H₂ gas mixture in an epitaxial reactor can be used to smooth the split surface as well. This technique has been shown to reduce the RMS surface roughness of the SOI film to less than 0.08 nm [75, 76].

Further Developments

The Smart-Cut® process has been subsequently used to transfer thin films of materials other than silicon. The literature reports the successful transfer of the following materials: SiC [77], GaAs, InP, and LiNbO₃ [78], SiGe [79], and other materials. The process can also be used to fabricate multilayer (Si/SiO₂/Si/SiO₂/Si/SiO₂/etc.) structures [80] and to transfer layers with patterned structures [81]. Debonding of the thin silicon overlayer from the handle wafer and its transfer to a third wafer using a polymer bonding technique has been demonstrated as well [82]. It has also been shown that, for silicon, co-implantation of boron with hydrogen permits one to perform the Smart-Cut® process with a lower thermal budget. This improvement has made it possible to transfer a thin silicon film onto quartz after annealing at a temperature of only 250°C [83].

8.2.6 Eltran®

The Eltran® (epitaxial layer transfer) technique combines the formation of a porous silicon layer, epitaxy, and wafer bonding to produce SOI wafers. The properties of porous silicon are discussed in the next section, followed by the description of the Eltran® fabrication process itself.

Porous Silicon

p-type silicon can readily be converted into porous silicon by electrochemical dissolution of p-type silicon in HF; the silicon wafer is immersed into an HF solution and a potential difference is applied between the sample and a platinum electrode dipped into the electrolyte. The degree of porosity of the layer can be controlled by adjusting the current utilized during the reaction. Porous

silicon was used in the 1980s to fabricate SOI wafers. The material was called “FIPOS”, which stands for “full isolation by porous silicon” [84, 85]. Porous silicon basically looks like a silicon “sponge”. It is full of wormhole-like cavities, but the silicon that has not been dissolved away by the electrochemical reaction is still single-crystal. In 1985 it was observed that high-quality silicon layers can be grown epitaxially on porous silicon [86]. Porous silicon has two key properties. Firstly, the surface pores can be filled up and sealed by baking the material in hydrogen. This property is used to prepare the surface before the growth of epitaxial silicon on porous silicon. Secondly, porous silicon has a an extremely large surface-to-volume ratio ($200\text{ m}^2/\text{cm}^3$) [87]. As a result, porous silicon has a very high chemical reactivity and it is possible to etch it in an $\text{HF}/\text{H}_2\text{O}_2$ solution with an extremely high selectivity (100 000:1) relative to silicon.

The Original Eltran® Process

The original Eltran® process was published in 1994 [88]. It comprises the following steps (Fig. 8.19): the formation of a blanket porous silicon layer on a silicon wafer (seed wafer), followed a hydrogen bake step to seal the surface pores, and the growth of a single-crystal epitaxial silicon film. After thermal growth of an oxide, the seed wafer is bonded to an oxidized wafer called the “handle wafer”. The bulk of the handle wafer is then removed by grinding and polishing until the porous silicon layer is reached. The porous silicon is removed by etching in an $\text{HF}/\text{H}_2\text{O}_2$ solution and a final H_2 annealing is applied to smooth the surface. One drawback of the original Eltran® process is that it requires two silicon wafers to produce a single SOI wafer.

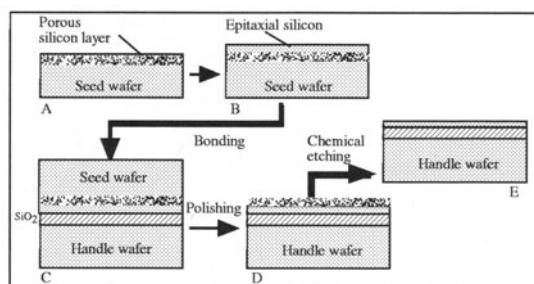


Fig. 8.19. The original Eltran® process. A, formation of a porous silicon layer; B, growth of epitaxial silicon; C, bonding to a handle wafer; D, polishing of the silicon wafer; E, porous-silicon etching

Second-Generation Eltran® Process

Like the Smart-Cut process, the second-generation Eltran® process uses $N + 1$ silicon wafers to produce N SOI wafers. The process is based on the formation of a double porous silicon layer, or, more exactly, a layer with two different porosities. The juxtaposition of these two layers generates mechanical stress near the interface between the two types of porous silicon. Then a high-pressure (20–60 MPa) water jet is used to “unzip” the seed wafer from the handle wafer along the stress region between the two porous silicon layers, thereby producing an SOI substrate and a recyclable seed wafer (Fig. 8.20). The important steps of the process are described next.

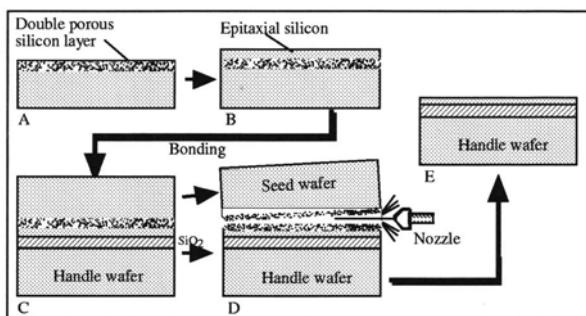


Fig. 8.20. Second-generation Eltran process. A, formation of a porous silicon layer; B, growth of epitaxial silicon; C, bonding to a handle wafer; D, splitting of porous silicon layer using a water jet; E, etching and H_2 annealing [90,91]

The porosity of a porous silicon layer can be modified by changing the current density used in the electrochemical reaction that produces it. The porosity increases from 20 to 65% when the current density is increased from 8 to 23 mA/cm² [89]; a porous layer with two different porosities can be created by changing the current density during porous-silicon formation. In the Eltran® process a low current density is used first to form a low-porosity layer at the surface of the seed wafer. A low-porosity surface layer is suitable for the subsequent growth of a high-quality epitaxial layer. The current density is then increased to produce a higher-porosity layer deeper underneath the surface. The difference in porosity generates mechanical stress between the two porous layers. A hydrogen bake step is applied to seal the pores at the surface, and a silicon layer (the future SOI layer) is grown by epitaxy. Hydrophilic bonding is used to attach the seed wafer to a handle wafer. Then, a jet of pressurized water with a diameter of 0.1 mm is directed at the edge of the wafer assembly. The water jet acts as a liquid wedge that splits the porous silicon at the region of maximum stress, i.e. where the two porosities meet. Because a liquid rather than a solid wedge is used, the splitting effect propa-

gates across the entire wafer assembly, and the porous silicon layer opens like a zipper. Once splitting has been achieved, the porous silicon layers can be removed from both the seed and the SOI wafer by etching in an HF/H₂O₂ solution. A final H₂ annealing is applied to smooth the surface.

8.3 Conclusion

After 20 years of intensive research and development efforts, SOI technology has reached maturity, and mass production of SOI wafers and SOI circuits has become an industrial reality. The number of SOI substrates available worldwide has increased from a few tens of thousands in the early 1990s to over 2 million in 2003. The available wafer sizes are 100, 200, and 300 mm. SOI devices offer tremendous advantages in terms of circuit speed and are particularly well adapted to the fabrication of low-voltage circuits for portable applications. Several major semiconductor manufacturers are now using SOI technology for important lines of products, such as microprocessors.

References

1. J.P. Colinge: *Silicon-on-Insulator Technology: Materials to VLSI* (Kluwer Academic, Dordrecht 1997)
2. S. Cristoloveanu, S.S. Li: *Electrical Characterization of Silicon-on-Insulator Materials and Devices* (Kluwer Academic, Dordrecht 1995)
3. J.B. Kuo, K.W. Su: *CMOS VLSI Engineering Silicon-on-Insulator (SOI)* (Kluwer Academic, Dordrecht 1998)
4. F. Ichikawa, Y. Nagatomo, Y. Kataura, S. Itoh, H. Matsuhashi, N. Hirashita, S. Baba: *Electrochemical Society Proceedings*, vol. 2003-05, p. 123 (2003)
5. S.B. Park, Y.W. Kim, Y.G. Ko, K.I. Kim, I.K. Kim, H.S. Kang, J.O. Yu, K.P. Suh: A 0.25- μ m, 600 Mhz, 1.5-V, fully depleted SOI CMOS 64-bit microprocessor. *IEEE Journal of Solid-State Circuits* **34**, 1436 (1999)
6. M. Canada, C. Akroul, D. Cawlthon, J. Corr, S. Geissler, R. Houle, P. Kartschoke, D. Kramer, P. McCormick, N. Rohrer, G. Salem, L. Warriner: A 580 Mhz RISC microprocessor in SOI. *Proceedings of the IEEE International Solid-State Conference* (1999) pp. 430-431
7. A.G. Aipperspach, D.H. Allen, D.T. Cox, N.V. Phan, S.N. Storino: A 0.2- μ m, 1.8 V, 550 Mhz, 64-b PowerPC microprocessor with copper interconnects. *IEEE Journal of Solid-State Circuits* **34**, 1430 (1999)
8. K. Shimomura, H. Shimano, N. Sakashita, F. Okuda, T. Oashi, Y. Yamaguchi, T. Eimori, M. Inuishi, K. Arimoto, S. Maegawa, Y. Inoue, S. Komori, K. Kyuma: A 1-V 46-ns 16-Mb SOI-DRAM with body control technique. *IEEE Journal of Solid-State Circuits* **32**, 1712 (1997)
9. T. Oashi, T. Eimori, F. Morishita, T. Iwamatsu, Y. Yamaguchi, F. Okuda, K. Shimomura, H. Shimano, S. Sakashita, K. Arimoto, Y. Inoue, S. Komori, M. Inuishi, T. Nishimura, H. Miyoshi: 16 Mb DRAM/SOI Technologies for Sub-1V Operation. *International Electron Devices Meeting* (1996) pp. 609-612

10. T. Eimori, T. Oashi, F. Morishita, T. Iwamatsu, Y. Yamaguchi, F. Okuda, K. Shimomura, H. Shimano, N. Sakashita, K. Arimoto, Y. Inoue, S. Komori, M. Inuishi, T. Nishimura, H. Miyoshi: Approaches to extra low voltage DRAM operation by SOI-DRAM. *IEEE Transactions on Electron Devices* **45**, 1000 (1998)
11. J.W. Park, Y.G. Kim, I.K. Kim, K.C. Park, K.C. Lee, T.S. Jung: Performance characteristics of SOI DRAM for low-power application. *IEEE Journal of Solid-State Circuits* **34**, 1446 (1999)
12. Y.H. Koh, M.R. Oh, J.W. Lee, J.W. Yang, W.C. Lee, C.K. Park, J.B. Park, Y.C. Heo, K.M. Rho, B.C. Lee, M.J. Chung, M. Huh, H.S. Kim, K.S. Choi, W.C. Lee, J.K. Lee, K.H. Ahn, K.W. Park, J.Y. Yang, H.K. Kim, D.H. Lee, I.S. Hwang: 1 Giga bit SOI DRAM with fully bulk compatible process and body-contacted SOI MOSFET structure. Technical Digest of the International Electron Devices Meeting (1997) pp. 579–582
13. Y.W. Kim, S.B. Park, Y.G. Ko, K.I. Kim, I.K. Kim, K.J. Bae, K.W. Lee, J.O. Y, U. Chung, K.P. Suh: Digest of Technical Papers of the IEEE International Solid-State Circuits Conference (1999) pp. 32–33
14. M. Itoh, Y. Kawai, S. Ito, K. Yokomizo, Y. Katakura, Y. Fukuda, F. Ichikawa: Electrochemical Society Proceedings, vol. 2001–3, p. 331 (2001)
15. T. Nishimura, Y. Inoue, K. Sugahara, S. Kusunoki, T. Kumamoto, S. Nakagawa, M. Nakaya, Y. Horiba, Y. Akasaka: Three dimension IC for high performance image signal processor. Technical Digest of the International Electron Devices Meeting (1987) pp. 111–114
16. J.P. Denton, G.W. Neudeck: Fully depleted dual-gated thin-film SOI P-MOSFETs fabricated in SOI islands with an isolated buried polysilicon back-gate. *IEEE Electron Device Letters* **17**, 509 (1996)
17. H.M. Manasevit, W.I. Simpson: Single-crystal silicon on a sapphire substrate. *Journal of Applied Physics* **35**, 1349 (1964)
18. S.S. Lau, S. Matteson, J.W. Mayer, P. Revesz, J. Gyulai, J. Roth, T.W. Sigmon, T. Cass: Improvement of crystalline quality of epitaxial Si layers by ion-implantation techniques. *Applied Physics Letters* **34**, 76 (1979)
19. J. Amano, K.A. Carey: Novel three-step process for low-defect-density silicon on sapphire. *Applied Physics Letters* **39**, 163 (1981)
20. P.K. Vasudev, D.C. Mayer: Characterization of CMOS devices in 0.5–1 μm silicon-on-sapphire films modified by solid phase epitaxy and regrowth (SPEAR). Comparison of Thin Film Transistor and SOI Technologies Symposium, pp. 35–39 (1984)
21. M.E. Roulet, P. Schwob, I. Golecki, M.A. Nicolet: Electrical properties of silicon-implanted furnace-annealed silicon-on-sapphire devices. *Electronics Letters* **15**, 527 (1979)
22. R. Reedy, J. Cable, D. Kelly, M. Stuber, F. Wright, G. Wu: UT_{Si} CMOS: A complete RF SOI solution. *Analog Integrated Circuits and Signal Processing*, vol. 25 no. 2, Nov. 2000, pp. 171–179
23. M. Megahed, M. Burgener, J. Cable, D. Staab, R. Reedy: UT_{Si} CMOS technology for system-on-chip solution. Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems. Digest of Papers (1998) pp. 94–99
24. K. Izumi, M. Doken, H. Ariyoshi: CMOS devices fabricated on buried SiO₂ layers formed by oxygen implantation into silicon. *Electronics Letters* **14**, 593 (1978)

25. G.F. Cerofolini, S. Bertoni, L. Meda, C. Spaggiari: Formation and stability of continuous buried SiO₂ layers in SIMOX. *Nucl. Instr. Methods Phys. Res. B* **84**, 234–237 (1993)
26. L. Meda, S. Bertoni, G.F. Cerofolini, C. Spaggiari: Structure and electrical characteristics of a thin buried oxide containing silicon inclusions. *Electrochemical Soc. Proceedings* **94-11**, 224–229 (1994)
27. S. Krause, M. Anc, P. Roitman: Evolution and future trends of SIMOX material. *MRS Bulletin* **23**, 25 (1988)
28. K. Izumi: History of SIMOX material. *MRS Bulletin* **23**, 20 (1998)
29. S. Nakashima, K. Izumi: Practical reduction of dislocation density in SIMOX wafers. *Electronics Letters* **26**, 1647–1649 (1990)
30. S. Nakashima, K. Izumi: SIMOX wafers with low dislocation density produced by a 100-mA-class high-current oxygen implanter. *Nuclear Instruments and Methods in Physics Research B* **55**, 847 (1991)
31. B. Aspar, C. Pudda, A.M. Papon, A.J. Auberton-Hervé, J.M. Lamure: Ultra thin buried oxide layers formed by low-dose SIMOX processes. *Electrochemical Society Proceedings* **94-11**, 62 (1994)
32. M.J. Anc, J.G. Blake, T. Nakai: Low-energy oxygen implantation for dose reduction in SIMOX. *Electrochemical Society Proceedings* **99-3**, 51–60 (1999)
33. M. Chen, X. Wang, Y. Dong, X. Liu, W. Yi, J. Chen, X. Wang: Formation of ultra-thin SOI by dose-energy optimization. *Proceedings of the IEEE International SOI Conference* (2002) pp. 113–114
34. A.J. Auberton-Hervé, B. Aspar, J.L. Pelloie: Low-dose SIMOX for ULSI applications. In: *Physical and Technical Problems of SOI Structures and Devices*, NATO ASI Series, High Technology 4, ed. by J.-P. Colinge, V.S. Lysenko, A.N. Nazarov (Kluwer Academic, Dordrecht 1995) pp. 3–14
35. S. Nakashima, T. Katayama, Y. Miyamura, A. Matsuzaki, M. Imai, K. Izumi, N. Ohwada: Thickness increment of buried oxide in a SIMOX wafer by high-temperature oxidation. *International SOI Conference Proceedings* (1994) pp. 71–72
36. M. Tachimori, S. Masui, T. Nakajima, K. Kawamura, I. Hamaguchi, T. Yano, Y. Nagatake: Quality innovation of SIMOX wafers by low-dose and high-temperature oxidation techniques. *Electrochemical Society Proceeding* **96-3**, 53 (1996)
37. A. Matsamura, K. Kawamura, T. Mizutani, S. Takayama, I. Hamaguchi, Y. Nagatake: Recent progress in low-dose SIMOX wafers fabricated with internal-thermal-Oxidation (ITOX) process. *Electrochemical Society Proceedings* **99-3**, 79–92 (1999)
38. D.K. Sadana: SOI for CMOS logic and memory applications. *Electrochemical Society Proceedings* **2001-2**, 474 (2001)
39. O.W. Holland, D. Fathy, D.K. Sadana: Formation of ultrathin, buried oxides in Si by O⁺ implantation. *Applied Physics Letters* **69**, 674 (1996)
40. L.P. Allen, M.L. Alles, R.P. Dolan, H.L. Hughes, P. McMarr: Thin BOX SIMOX silicon-on-insulator structures for radiation tolerant advanced electronics. *Microelectronic Engineering* **36**, 383 (1997)
41. J. Blake, K. Dempsey, R. Dolan, Y. Erokhin, P. Powell, S. Richards: 300 mm ultra-thin AdvantoxTM MLD SIMOX wafers manufactured using i2000 oxygen implanter. *Proceedings of the IEEE International SOI Conference* (2002) pp. 109–110

42. A. Ogura: Novel SIMOX with BOX at damage peak. *Electrochemical Society Proceedings* **99-3**, 61 (1999)
43. A. Ogura: Novel SOI fabrication process by light ion implantation and annealing in oxygen including atmosphere. *Extended Abstracts of the International Conference on Solid-State Devices and Materials* (2001) pp. 240–241
44. R. Stengl, T. Tan, U. Gösele: A model for the silicon wafer bonding process. *Japanese Journal of Applied Physics* **28**, 1735 (1989)
45. W.P. Maszara, G. Goetz, A. Cavigilia, J.B. McKitterick: Bonding of silicon wafers for silicon-on-insulator. *Journal of Applied Physics* **64**, 4943 (1988)
46. Q.Y. Tong, U. Gösele: *Semiconductor Wafer Bonding Science and Technology*, The Electrochemical Society Series (Wiley, New York 1999)
47. P. Amirfeiz, S. Bengtsson, M. Bergh, E. Zanghellini, L. Börjesson: Formation of silicon structures by plasma activated binding. *Electrochemical Society Proceedings* **99-35**, 29 (1999)
48. C.A. Desmond-Colinge, U. Gösele: Wafer-bonding and thinning technologies. *MRS Bulletin* **23**, 30 (1988)
49. S.D. Collins: Etch stop techniques for micromachining. *Journal of the Electrochemical Society* **144**, 2242 (1997)
50. M. Bruel: Silicon on insulator material technology. *Electronics Letters* **31**, 1201 (1995)
51. M. Bruel: The history, physics, and applications of the Smart-Cut® process. *MRS Bulletin* **23**, 35 (1998)
52. M. Kaminsky: Plasma contamination and wall erosion in thermonuclear reactors. *IEEE Transactions on Nuclear Science* **18**, 208 (1971)
53. K. Kamada, Y. Kazumata, K. Kubo: Observation of blistering and amorphization on germanium after 450 keV Ar+ ion bombardment. *Radiation Effects* **28**, 43 (1976)
54. F.F. Komarov, V.S. Solov'yev, S.Y. Shiryayev: Crystallographic nature and formation mechanisms of highly irregular structures in implanted and annealed Si layers. *Radiation Effects* **42**, 169 (1979)
55. C. Ascheron, H. Bartsch, A. Setzer, A. Schindler, P. Paufler: The effect of hydrogen implantation induced stress on GaP single crystals. *Nuclear Instruments and Methods in Physics Research B* **28**, 350 (1987)
56. J.H. Evans: An interbubble fracture mechanism of blister formation on helium-irradiated metals. *Journal of Nuclear Materials* **68**, 129 (1977)
57. H. Van Swijgenhoven, L.M. Stals, G. Knuyt: Helium bubble and blister formation for nickel and amorphous Fe–Ni–Mo–B alloy during 5 keV He+ irradiation at temperatures between 200 K and 600 K. *Nuclear Instruments and Methods* **209/210**, 461 (1983)
58. K. Ono, T. Kino, K. Kamada, H. Osono: Orientation dependence of flaking of ion irradiated aluminum single crystals. *Japanese Journal of Applied Physics* **25**, 1475 (1986)
59. K. Wittmaack, P. Blank, W. Wach: High fluence retention of noble gases implanted in silicon. *Radiation Effects* **39**, 81 (1978)
60. E. Ligeon, A. Guivarc'h: Hydrogen implantation in silicon between 1.5 and 60 keV. *Radiation Effects* **27**, 129 (1976)
61. M. Bruel: Process for the production of thin semiconductor material films. U.S. Patent 5,374,564 (1994)

62. C.F. Cerofolini, L. Meda, R. Balboni, F. Corni, S. Frabboni, G. Ottaviani, R. Tonini, M. Anderle, R. Canteri: Hydrogen-related complexes as the stressing species in high-fluence, hydrogen-implanted, single-crystal silicon. *Physical Review B* **46**, 2061 (1992)
63. M. Bruel: The history, physics, and applications of the Smart-Cut® process. *MRS Bulletin* **23**, 35 (1998)
64. J. Grisolia, G. Ben Assayag, A. Claverie, B. Aspar, C. Lagahe, L. Laanab: A transmission electron microscopy quantitative study of the growth kinetics of H platelets in Si. *Applied Physics Letters* **76**, 852 (2000)
65. F.A. Reboreda, M. Ferconi, S.T. Pantelides: Theory of nucleation, growth, and structure of hydrogen-induced extended defects in silicon. *Physical Review Letters* **82**, 4870 (1999)
66. M.K. Weldon, V.E. Marsico, Y.J. Chabal, A. Agarwal, D.J. Eaglesham, J. Sapjeta, W.L. Brown, D.C. Jacobson, Y. Caudano, S.B. Christman, E.E. Chaban: On the mechanism of the hydrogen-induced exfoliation of silicon. *Journal of Vacuum Science and Technology B* **15**, 1065 (1997)
67. T.H. Lee, Q.Y. Tong, Y.L. Chao, L.J. Huang, U. Gösele: Silicon on quartz by a smarter cut process.. *Electrochemical Society Proceedings* **97-23**, 27 (1997)
68. B. Aspar, C. Lagahe, H. Moriceau, E. Jalaguier, A. Mas, O. Rayssac, A. Soubie, B. Biasse, M. Bruel: The Smart-Cut® process: status and development. *Electrochemical Society Proceedings* **99-35**, 48 (1999)
69. K. Hentinnen, I. Suni, S.S. Lau: *Appl. Phys. Lett.* **76**, 2370 (2000)
70. B. Aspar, H. Moriceau, E. Jalaguier, C. Lagahe, A. Soubie, B. Biasse, A.M. Papon, A. Claverie, J. Grisolia, G. Benassayag, F. Letertre, O. Rayssac, T. Barge, C. Maleville, B. Chyselen: *Journal of Electronic Materials* **30**, 834 (2001)
71. Q.Y. Tong, R.W. Bower: Beyond “Smart-Cut®”: recent advances in layer transfer for material integration. *MRS Bulletin* **23**, 40 (1998)
72. K. Hentinnen, I. Suni, S.S. Lau: Mechanically induced Si layer transfer in hydrogen-implanted Si wafers. *Applied Physics Letters* **76**, 2370 (2000)
73. W.G. En, I.J. Malik, M.A. Bryan, S. Farrens, F.J. Henley, N.W. Cheung, C. Chan: The Genesis process: a new SOI wafer fabrication method. *Proceedings of the IEEE International SOI Conference* (1998) pp. 163–164
74. J.T.S. Lin, J. Peng, T.H. Lee: NovaCut™ process: fabrication of silicon on insulator materials. *Proceedings of the IEEE International SOI Conference* (2002) pp. 189–190
75. A. Thilderkvist, S. Kang, M. Fuerfanger, I. Malik: Surface finishing of cleaved SOI films using epi technologies. *Proceedings of the IEEE International SOI Conference* (2000) pp. 12–13
76. M.I. Current, I.J. Malik, M. Fuerfanger, A. Flat, J. Sullivan, S. Kang, H.R. Kirk, M. Norcott, D. Teoh, P. Ong, F.J. Henley: Surface roughness and device layer thickness for ultra-thin SOI. *Proceedings of the IEEE International SOI Conference* (2002) pp. 111–112
77. J.-P. Joly, B. Aspar, M. Bruel, L. Di Cioccio, F. Letertre, E. Hugonnard-Bruere: New SiC on insulator wafers based on the Smart-Cut® approach and their potential applications. In: *Progress in SOI Structures and Device Operating at Extreme Conditions*. *Proceedings of the Nato Advanced Research Workshop*, ed. by F. Balestra, A.N. Nazarov, V.S. Lysenko (Kluwer Academic, Dordrecht, Netherlands 2002) pp. 31–38

78. B. Aspar, C. Lagahe, H. Moriceau, E. Jalaguier, A. Mas, O. Rayssac, A. Soubie, B. Biasse, M. Bruel: The Smart-Cut® process: status and development. *Electrochemical Society Proceedings* **99-35**, 48 (1999)
79. L.J. Huang, J.O. Chu, D.F. Canaperi, C.P. D'Emic, R.M. Anderson, S.J. Koester, H.S.P. Wong: SiGe-on-insulator prepared by wafer bonding and layer transfer for high-performance field-effect transistors. *Applied Physics Letters* **78**, 1267 (2001)
80. C. Maleville, T. Barge, B. Ghyselen, A.J. Auberton: Multiple SOI layers by multiple Smart-Cut® transfers. *Proceedings of the IEEE International SOI Conference* (2000) pp. 134–135
81. B. Aspar, M. Bruel, M. Zussy, A.M. Cartier: Transfer of structured and patterned thin silicon films using the Smart-Cut® process. *Electronics Letters* **32**, 1985 (1996)
82. C. Colinge, B. Roberds, B. Doyle: Silicon layer transfer using wafer bonding and debonding. *Journal of Electronic Materials* **30**, 841 (2001)
83. T.H. Lee, Q.Y. Tong, Y.L. Chao, L.J. Huang, U. Gosele: Silicon on quartz by a smarter cut process. *Electrochemical Society Proceedings* **97**, 27 (1997)
84. K. Imai: A new dielectric isolation method using porous silicon. *Solid-State Electronics* **24**, 159 (1981)
85. S.S. Tsao: Porous silicon techniques for SOI structures. *IEEE Circuits and Devices Magazine* **3**, 3 (1987)
86. M.I.J. Beale, N.G. Chew, A.G. Cullis, D.B. Gasson, R.W. Hardeman, D.J. Robbins, I.M. Young: A study of silicon MBE on porous silicon substrates. *Journal of Vacuum Science and Technology B* **3**, 732 (1985)
87. T. Yonehara, K. Sakaguchi, N. Sato: Eltran®: epitaxial layer transfer. *Electrochemical Society Proceedings* **99-3**, 111 (1999)
88. T. Yonehara, K. Sakaguchi, N. Sato: Epitaxial layer transfer by bond and etch back of porous Si. *Applied Physics Letters* **64**, 2108 (1994)
89. T. Yonehara, K. Sakaguchi: Eltran® (SOI-epi wafer™) technology. In: *Progress in SOI Structures and Device Operating at Extreme Conditions*. *Proceedings of the Nato Advanced Research Workshop*, ed. by F. Balestra, A.N. Nazarov, V.S. Lysenko (Kluwer Academic, Dordrecht, Netherlands 2002) pp. 39–86
90. K. Sakaguchi, T. Yonehara: SOI wafers based on epitaxial technology. *Solid-State Technology* **43-6**, 88 (2000)
91. K. Sakaguchi, K. Yanagita, H. Kurisu, H. Suzuki, K. Ohmi, T. Yonehara: Eltran®: by splitting porous Si layers. *Electrochemical Society Proceedings* **99-3**, 117 (1999)

Part IV

Lattice Defects

9 Defect Spectroscopy

H.G. Grimmeiss

9.1 Introduction

The impressive progress in semiconductor electronics during recent decades can be traced to a unique combination of basic conceptual advances, the perfection of new materials and the development of new device principles. Ever since the beginning, we have witnessed a fantastic growth in silicon technology. This development would hardly have been possible without an increased understanding of semiconductor materials and a better insight into the important role of defects for improving material and device quality. It is therefore not surprising that a large variety of measurement techniques for the characterisation and identification of defects has been developed right from the beginning, with the aim of further increasing our understanding of defects.

In contrast to compound semiconductors, which are in general tainted with non-stoichiometry and other intrinsic imperfections, defect studies are quite often much easier to perform in elementary semiconductors such as silicon and germanium. This does not necessarily imply that the electronic structures of defects in elementary semiconductors are less complex than in compound semiconductors. Several studies on silicon have shown that even the electronic structures of point defects can be rather complex. Compared with group III and V impurities, which have rather small binding energies and are, hence, well described by the effective-mass approximation [1], it is well known that impurities from other groups of the periodic table in most cases exhibit very different properties. This is particularly true for transition metals.

The good understanding of group III and V impurities in silicon in the early days was at least partly due to the fact that their electronic structure is easily studied by absorption and/or photoconductivity, and that the corresponding spectra in general are line spectra, which offer a great amount of information. Although comparably good line spectra of group VI impurities with considerably larger binding energies were already available in 1962 owing to the pioneering work of Krag and Zeiger [2], it was the introduction of junction space charge techniques that delayed the wider appreciation of characterisation techniques involving high-resolution spectroscopy. Those latter techniques were rediscovered at the end of the 1970s and contributed to several breakthroughs, in particular with respect to the understanding of deep defects.

A large variety of measurement techniques for the characterisation and identification of defects in silicon are available. These techniques comprise both bulk and junction space charge techniques (JSCTs). Earlier studies involved mostly bulk measurements such as photoconductivity, electrical and magnetic methods, as well as absorption and various forms of luminescence measurements. All these methods allowed the study of thermal "activation energies" and/or optical threshold energies, however, only very rarely could absolute values of fundamental electronic parameters be determined.

The situation changed markedly with the introduction of junction space charge techniques, such as photocurrent [3] and dark capacitance [4] measurement techniques. Most of these methods, in particular deep-level transient spectroscopy (DLTS) [5], allow the determination of defect concentrations and the direct measurement of important electronic parameters such as thermal and optical emission and capture rates.

In parallel with the development of these characterisation techniques, the application of spin-dependent measurements such as electron paramagnetic resonance (EPR) [6] provided another breakthrough, this time in the area of chemical identification. Unfortunately, the data obtained with these techniques can very rarely be directly related to defects that have been studied by, for example, JSCTs. This implies that many defects in semiconductors that have been accurately characterised (for example, certain gold centres in silicon) have never been identified, and vice versa, since spin-dependent measurement techniques are in general not the perfect tools for studying electronic parameters. Techniques combining spin-dependent data with other methods [7] such as characterisation measurements are therefore extremely important. Photo-EPR is one such example.

Both bulk measurement techniques and JSCTs have been developed further and, in particular, high-resolution absorption and photo-thermal ionisation spectroscopy (PTIS), as well as admittance spectroscopy techniques, have been applied to study defects in silicon and in other semiconductors. By using these high-resolution measurement techniques, information is readily obtained on binding energies, charge states, symmetry and lattice relaxation as well as on the kind of transition (donor- or acceptor-like) and whether the defect is isolated or a complex. In contrast to luminescence measurements, both PTIS and high-resolution absorption measurements can be used in the study of defects, regardless of whether they are considered as radiative or non-radiative defects.

Additional information on the energy structure and electronic properties of defects in silicon is often obtained by applying perturbation spectroscopy, such as Zeeman spectroscopy [8] and piezo-spectroscopy [9]. In particular the latter technique has been proven to be very useful for the study of defects in silicon.

9.2 Fundamental Parameters Characterising the Properties of Defects

The electronic properties and thus the influence of defects on the electrical and optical properties of a semiconductor are usually described by the binding energy of the ground state E_T , the emission rates $e_{n,p}$ and the capture constants $c_{n,p}$. Here, $e_n = e_n^o + e_n^t$ and $e_p = e_p^o + e_p^t$ are the sums of the optical and thermal emission rates of electrons and holes, respectively. Similar notation is used for the capture constants. Since for all kinetic processes the total concentration of the defect N_{TT} is of great importance, one of the goals of defect spectroscopy is to determine the ten fundamental parameters E_T , $e_{n,p}^o$, $e_{n,p}^t$, $c_{n,p}^o$, $c_{n,p}^t$ and N_{TT} as accurately as possible.

Thermal and optical emission processes are separated by performing measurements in darkness or under illumination below the freeze-out temperature. Most JSCTs for the measurement of capture constants are methods for studying the sum of the capture constants of the dominant radiative and non-radiative capture processes, $c_{n,p} = c_{n,p}^t + c_{n,p}^o$.

The capture constant $c_{n,p}$ is related to the thermal emission rate $e_{n,p}^t$ by the detailed balance relationship (Engström and Alm 1978) [10]:

$$e_{n,p}^t = c_{n,p} N_{c,v} \exp(-\Delta G_{n,p}/kT). \quad (9.1)$$

Here, N_c or N_v is the effective density of states in the conduction or valence band, respectively, and ΔG_n or ΔG_p is the change in the Gibbs free energy needed to emit an electron or hole from the defect. The optical binding energy $\Delta G_{n,p}^o$ of the defect with respect to one of the energy bands can be obtained from the Gibbs free energy using the relation (Rees and Almblad 1980, unpublished)

$$\Delta G_{n,p}^o = \Delta G_{n,p} + kT \ln g, \quad (9.2)$$

provided the electronic degeneracy g is known. Except where otherwise mentioned, the conventional notation for the energy position of a centre E_T (which in the above notation corresponds to $\Delta G_p^o = E_g - \Delta G_n^o$) is used in the following, where E_g is the energy bandgap. Since both $\Delta G_{n,p}$ and $c_{n,p}$ are usually temperature dependent, the energy position of a defect cannot be determined by measuring the temperature dependence of the thermal emission rate alone (for example by DLTS). $\Delta G_{n,p}$ and hence $\Delta G_{n,p}^o$ are only obtained when the temperature dependences of both $e_{n,p}^t$ and $c_{n,p}$ are known.

This is readily shown by taking into account that $\Delta G_{n,p}$ is related to both the change in enthalpy $\Delta H_{n,p}$ and the total change in entropy $\Delta S_{n,p}$ by the relation

$$\Delta G_{n,p} = \Delta H_{n,p} - T \Delta S_{n,p}. \quad (9.3)$$

Assuming an exponential temperature dependence of the capture cross section (which in many cases is a good approximation),

$$\sigma_{n,p} = c_{n,p}/v_{th} = \sigma_o \exp(-\Delta H_{n,p}^c/kT), \quad (9.4)$$

the thermal emission rate per centre is given by

$$e_{n,p}^t/T^2 = M' \exp(\Delta S_{n,p}/k) \exp[-(\Delta H_{n,p} + \Delta H_{n,p}^c)/kT]. \quad (9.5)$$

Here, v_{th} is the thermal velocity, σ_o is a temperature-independent pre-factor and $\Delta H_{n,p}^c$ is an activation energy which can be interpreted as the change in enthalpy due to the capture of a charge carrier. $M' = \sigma_o v_{th} N_{c,v} T^{-2}$ is then a temperature-independent factor. Plotting $\log(e_{n,p}^t/T^2)$ versus $1/T$ gives an activation energy

$$\Delta E' = \Delta H_{n,p} + \Delta H_{n,p}^c \quad (9.6)$$

that in most cases is quite different from $\Delta G_{n,p}$ and, hence, does not give any information on the energy position of the defect, in contrast to what is often pretended. It is also often forgotten that $\sigma_{n,p}$ is not accessible from such a plot, since the pre-exponential factor M' contains σ_o and not $\sigma_{n,p}$. Also, for vanishing $\Delta H_{n,p}^c$, capture cross-sections cannot be determined from a plot of $\log(e_{n,p}^t/T^2)$ versus $1/T$ if $\Delta S_{n,p}$ is not known, which is often the case. Considering that $\Delta S_{n,p}$ is difficult to obtain from measurement technique others than JSCTs, it is quite obvious that capture cross-sections can only be obtained from separate measurements.

Using (9.4) and (9.5), it is readily seen that $\Delta H_{n,p}$ can be calculated directly from the temperature dependence of the emission rate and the capture cross-section without knowing the absolute values of those quantities. $\Delta G_{n,p}$ is obtained by inserting absolute values of $e_{n,p}^t$ and $c_{n,p}$ into (9.1). Considering that absolute values of capture cross-sections are in general difficult to measure, it is recommended that one should perform this procedure for several temperatures and plot the calculated $\Delta G_{n,p}$ values versus T . From the slope of the regression line, the absolute value of $\Delta S_{n,p}$ is obtained, and at $T = 0$, the $\Delta G_{n,p}$ value gives $\Delta H_{n,p}$. This approach both is simple and has the advantage of presenting $\Delta G_{n,p}$ as a function of temperature; $\Delta G_{n,p}$ is the proper energy to be used for the comparison of thermal data with optical data, at least at lower temperatures.

It has already been mentioned that absolute values of capture cross-sections are difficult to obtain for certain defects. From the above procedure for calculating thermodynamic quantities, it is not possible to check whether or not the measured cross-sections are correct or whether they differ by a constant factor. If the cross-sections differ by a constant factor, the change in entropy but not the change in enthalpy will vary, as long as the temperature dependence of $e_{n,p}^t$ and $c_{n,p}$ has been measured properly. Incorrect absolute values of $c_{n,p}$ are therefore not seldom the explanation for unrealistic values of changes in entropy in the literature.

As pointed out earlier, the capture processes of charge carriers via a defect may differ depending on the electronic structure of the defect. For example, considering that the energy structure of a defect in general consists of a ladder of excited states, the assumptions made in (9.4) may not always be fulfilled. It has therefore been suggested that the capture of charge carriers, at least

at the beginning of the capture process, is sometimes better described by a cascade process, as proposed by Lax [11] and revised by Abakumov et al. [12]. Using this model, the capture cross-section of electrons is then given by

$$\sigma_n^t = (2\Gamma k/m_n^* s)T^{-3}, \quad (9.7)$$

where s is the velocity of sound, Γ a temperature-independent factor and m_n^* the density-of-states effective mass.

Another example is the model of Gibb et al. [13], who proposed carrier capture into the ground state via a single excited state in competition with thermal re-emission back to the valence or conduction band.

Several studies performed with different assumptions about the temperature dependence of the capture cross-section have shown that the change in enthalpy obtained from a combined analysis of σ_n^t and e_n^t is independent of the assumed temperature dependence of σ_n^t as long as a proper analysis of e_n^t is performed and (9.1) is not violated. A more comprehensive study of the temperature dependence of capture cross-sections may nevertheless be useful, since different capture processes give rise to different temperature dependences, which therefore in turn give information on the capture process.

9.3 Junction Space Charge Techniques

During recent decades, a great variety of different junction space charge techniques have been developed for the study of electronic properties of defects. Owing to the limited space available, only the common basic principles of these techniques will be discussed.

Junction space charge techniques are very useful and, in principle, simple techniques for a fast determination of important parameters such as concentrations, emission rates, capture constants and thermal activation energies. However, most JSCTs do not exhibit sufficient energy resolution for a comprehensive analysis of, for example, the energy structures of defects. Independent of the labelling of some of the JSCTs, they are in general not spectroscopic methods, as shown later when high-resolution absorption and other measurement techniques are discussed.

9.3.1 Capacitance Techniques

Both steady-state and transient capacitance techniques are popular techniques for the characterisation of defects. Transient capacitance measurements can be performed either by keeping the reverse voltage constant and measuring the change in capacitance or by keeping the capacitance of the diode constant and measuring the change in reverse voltage [14]. In the latter case, the degree of defect compensation is far less critical than in the

former case but since this technique is not used as often as the former one, the discussion here is restricted to methods with constant reverse bias.

The transient dark capacitance technique was originally suggested by Williams [15]. A comprehensive treatment of this measurement method was later given by Sah et al. [4]. Lang [5] extended this technique to a high-frequency transient thermal scanning method (DLTS), which is a fast characterisation method and not limited to majority carrier traps.

Thermal Measurement Techniques

When transient capacitance techniques are used, the reverse-bias voltage V_R of a Schottky barrier or one-sided p–n junction is momentarily reduced for a short time t_s . The depletion region then contracts, thereby making free charge carriers available for recombination processes. The spatial location in which first recombination and then emission take place is the region through which the depletion layer is moved.

To be more specific, the following discussion is related to abrupt p⁺–n junctions with a single donor-like defect of concentration N_{TT} in the upper half of the bandgap. The capacitance of such a junction is given by

$$C(t) = A\varepsilon\varepsilon_0/W(t) = [A^2q\varepsilon\varepsilon_0N_I(t)/2(V_D + V_R)]^{1/2}, \quad (9.8)$$

where V_D is the diffusion voltage, A is the area of the junction, ε is the dielectric constant and W is the width of the space charge region. The above relation shows that the capacitance of such a junction depends on the space charge and, hence, on the spatial variation and the total concentration of all ionised impurities $N_I(t)$, which in our case is given by

$$N_I(t) = N_D + p_T(t) = N_D + N_{TT} - n_T(t). \quad (9.9)$$

Here N_D is the total concentration of shallow donors, which, owing to their small binding energies, are all assumed to be ionised, and $p_T(t)$ and $n_T(t)$ are the concentrations of empty and filled defect centres, respectively. For diodes with $N_D \gg N_{TT}$, expansion of (9.8) in a Taylor series gives the following relation for the total change of the diode capacitance:

$$\begin{aligned} \Delta C &= C[n_T(t_2)] - C[n_T(t_1)] \\ &= [n_T(t_1) - n_T(t_2)](1/2)[A^2q\varepsilon\varepsilon_0 t_2 2(V_D + V_R)N_D]^{1/2} \\ &= \text{const.}[n_T(t_1) - n_T(t_2)], \end{aligned} \quad (9.10)$$

where $n_T(t)$ is the concentration of defects occupied by electrons at time t . The constant of (9.10) can either be determined experimentally or be calculated if V_D and N_D are known [16]. With properly chosen initial and final conditions, the total concentration of the defect N_{TT} is then readily obtained from (9.10). This is, for example, the case when $e_p^t \ll e_n^t$ and the reverse-biased diode is short-circuited in darkness for a short time interval so that

$n_T(t_1) = n_T(0) = N_{TT}$ and $n_T(t_2) = n_T(\infty) = 0$. This technique can also be used for profiling N_{TT} by performing the measurements at different reverse biases.

The change in the electron occupancy of a defect is given by the following rate equation:

$$dn_T/dt = dp/dt - dn/dt = e_p p_T - c_p p n_T - e_n n_T + c_n n p_T. \quad (9.11)$$

Within the space charge region, where both n and p are negligibly small, integration of (9.11) gives

$$\begin{aligned} n_T(t) &= N_{TT}(e_p^0 + e_p^t)/(e_p^0 + e_p^t + e_n^0 + e_n^t) \\ &+ [n_T(0) - N_{TT}(e_p^0 + e_p^t)/(e_p^0 + e_p^t + e_n^0 + e_n^t)] \\ &\exp [-(e_p^0 + e_p^t + e_n^0 + e_n^t)t]. \end{aligned} \quad (9.12)$$

If, after short-circuiting the diode, not only ΔC but also the time constant τ of the capacitance transient is measured in darkness ($e_p^0 = e_n^0 = 0$), one obtains

$$\tau = 1/(e_p^t + e_n^t). \quad (9.13)$$

For $e_p^t \ll e_n^t$, the absolute value of the thermal emission rate for electrons is then directly obtained from the decay time constant τ .

When e_n^t and e_p^t are of comparable magnitude, absolute values of e_n^t are obtained by using both ΔC and τ . Considering that in this case $n_T(t_1) = N_{TT}$ and $n_T(t_2 = \infty) = N_{TT} e_p^t/(e_p^t + e_n^t)$, one obtains from (9.10), (9.12) and (9.13)

$$e_n^t = \Delta C(\text{const. } N_{TT} \tau)^{-1}. \quad (9.14)$$

Because the constant is only weakly temperature-dependent, at least in the temperature range where the shallow donors are completely ionised, the temperature dependence of e_n^t is in most cases readily obtained by using (9.14). Once e_n^t is known, e_p^t can be determined using (9.13).

The measuring technique discussed above can also be used for measuring capture constants if the measurements are performed in darkness ($e_p^0 = e_n^0 = 0$) and below the freeze-out temperature ($e_p^t = e_n^t = 0$). When the reverse bias is reduced, the depletion region contracts and free carriers enter the region through which the depletion region has moved. This process is very fast and results in an electrically neutral region almost simultaneously with the presence of free charge carriers available for capture into empty defect states. The concentration of free charge carriers in the region where the capture process occurs is the same as in the neutral region. For centres partly empty before the reverse bias is reduced ($n_T(0) < N_{TT}$) (obtained, for example, by illuminating the sample with sub-bandgap light), the time dependence of $n_T(t)$ is obtained by integrating (9.11) ($p = 0$), giving

$$n_T(t) = N_{TT} - [N_{TT} - n_T(0)] \exp (c_n n t). \quad (9.15)$$

Inserting (9.15) into (9.10) shows that the time constant for the corresponding change in capacitance is the same as for $n_T(t)$, i.e.

$$\tau = 1/c_n^n. \quad (9.16)$$

The change in the capacitance due to the reduction of the reverse bias is measured after the diode is reverse biased again (see Fig. 9.1). Since $c_n^t = \sigma_n^t v_{th}$, the thermal capture cross-section σ_n^t for a non-radiative capture process can be obtained by dividing the capture constant by the average thermal velocity of electrons $v_{th} = (3kT/m_n^*)^{1/2}$. It is easily shown that this measurement technique is not limited to defects in the upper half of the bandgap.

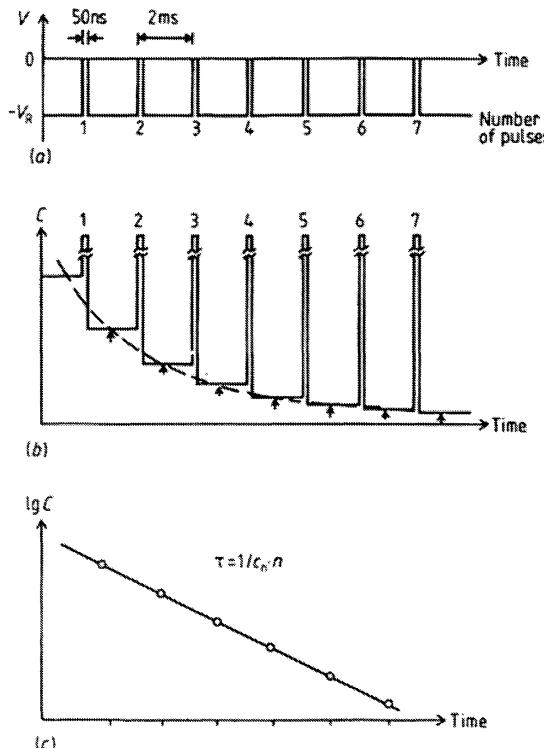


Fig. 9.1. Diagram illustrating measurements of capture cross-sections at temperatures below the freeze-out temperature (from [16])

Once the temperature dependence and the absolute values of c_n and e_n^t are known, ΔH_n is obtained from (9.4) and (9.5), ΔG_n is obtained from (9.1), ΔS_n is obtained from (9.3), and ΔG_n^o is obtained by using (9.2).

The single-transient capacitance techniques for measuring capture and emission constants have been developed further into methods that repetitively

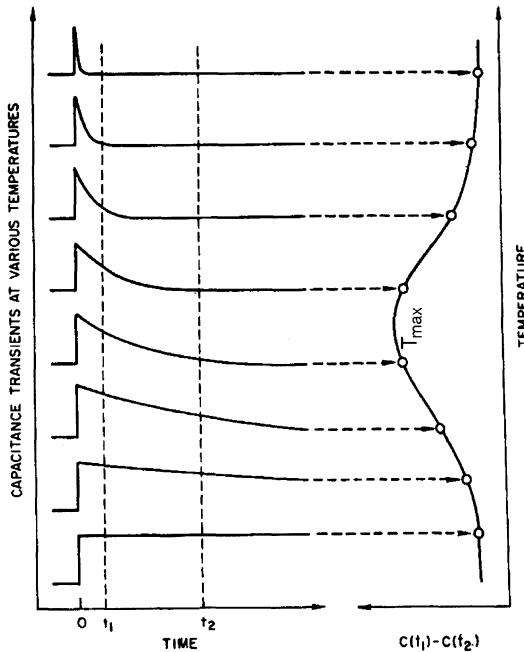


Fig. 9.2. Schematic representation of DLTS measurements (from [17])

apply or reduce the reverse bias of the diode [17]. These techniques simplify the initial characterisation of semiconductors, in particular when more than one defect is present. This type of technique, which originally was called deep-level transient spectroscopy (DLTS) [5], makes use of the fact that under certain experimental conditions, the capacitance transient, when the reverse-biased diode is pulsed to lower reverse bias or into forward bias, is exponential, with an initial amplitude that is proportional to the defect concentration (see (9.10)) and a time constant τ given, for example, by (9.13). By measuring the capacitance transient at two different but fixed times t_1 and t_2 (rate window) by using, for example, a double boxcar averager or lock-in amplifier, it can be shown that the difference of the capacitances $\Delta C = C(t_1) - C(t_2)$, when measured as a function of temperature, goes through a maximum at a temperature T_{\max} (Fig. 9.2). For a $p^+ - n$ junction and a defect in the upper half of the bandgap ($e_p^t \ll e_n^t$), the time constant of the transient and the emission rate at T_{\max} are correlated with the rate window by the relation

$$e_n = 1/\tau = \ln(t_2/t_1)/(t_2 - t_1). \quad (9.17)$$

The thermal activation energy ΔE^t of the defect is obtained by repeating the temperature scan for different rate windows t_1 , t_2 and plotting $\log(e_n/T^2)$ versus $1/T$.

If there are several defects present, the transient will be a sum of exponentials, each having a time constant at some characteristic temperature equal to the value set by the rate window. The rate window separates the exponentials so that scanning the temperature produces a series of peaks (one for each defect contributing to the measured transient), whose amplitudes are proportional to the corresponding defect concentrations.

DLTS techniques may also be used for studying capture constants. Reverse biasing a $p^+ - n$ junction with a defect in the upper half of the band gap, for example, will fill empty defect states at a rate given by (9.15). Hence, using a constant rate window but varying the filling-pulse width $t_s \ll 1/c_n$ will result in DLTS peaks of varying height. Plotting the change of the DLTS peak height logarithmically against t_s will result in a straight line, the slope of which corresponds to the thermal capture constant of electrons at T_{\max} . Performing these measurements at various rate windows leads to $c_n(T)$.

A large number of refinements and variations of the original DLTS method has been developed. One of the most complete lists of references on this subject is given in [42].

Though JSCTs are widely used, their applicability is sometimes restricted owing to inherent limitations given by the insufficient energy resolution. JSCTs are therefore in general not spectroscopic measurement techniques. Other limitations of these techniques become evident when one is studying, for example, emission processes. Using JSCTs implies that emission processes are always studied in the presence of high electric fields, whereas capture processes are measured under zero-field conditions. It has been shown that electric fields affect activation energies, for example. Electric-field effects on the thermal emission of charge carriers may in principle arise from different physical processes and hence result in different field dependences. It has therefore been suggested that a study of the electric-field-assisted thermal emission may be used to distinguish between donor centres and acceptor centres [18] and/or establish the charge state of a centre [19]. Detailed studies of field-enhanced emission processes on single substitutional chalcogen donors in silicon have shown that these distinctions are in general not possible with JSCTs owing to their complexity [20]. On the other hand, both the charge state of defects and the distinction between donor-like and acceptor-like transitions are readily studied by using high-resolution optical measurements, as shown below.

Field-enhanced thermal emission processes can nevertheless be investigated by JSCTs, for example by using single-shot measurements [20]. In this case, a narrow range within the space charge region of the junction is selected by measuring the difference between two transients that are recorded at two slightly different pulse voltages at constant reverse bias. Under these circumstances, thermal emission processes are studied at a rather constant electric field strength. The field strength is varied when the reverse bias is changed. Using this measurement technique, good agreement has been obtained be-

tween the zero-field enthalpies ΔH_n and the optical binding energies at low temperatures in spite of rather large field effects.

Another limitation of JSCTs arises from the RC product of the diode. Knowledge of the influence of the RC product is of particular interest when one is measuring large capture constants, since the upper limit of the capture cross-section that can be measured by DLTS is normally determined by the RC product of the diode. Further complications arise from the fact that the width of the space charge region changes during the applied capture pulse and that the concentration of captured electrons is therefore not constant, but will vary depending on the position within the space charge region of the reverse-biased diode. It has been shown [21] that the value of the RC product must be at least a factor of ten smaller than the time constant of the capture constant in order to avoid considerable errors in the determination of the capture constant. These errors can easily remain undetected in an ordinary analysis.

Optical Measurement Techniques

Transient capacitance measurements have also been applied for measuring optical emission rates, by keeping either the reverse bias or the diode capacitance constant. In what follows, the measurements are assumed to be performed at constant reverse bias below the freeze-out temperature of the defect ($e_p^t = e_n^t = 0$). Equation (9.12) then gives

$$n_T(t) = N_{TT}e_p^o/(e_p^o + e_n^o) + [n_T(0) - N_{TT}e_p^o/(e_p^o + e_n^o)] \exp[-(e_p^o + e_n^o)t], \quad (9.18)$$

implying a similar time constant for ΔC . If the initial conditions are chosen such that $n_T(0) = N_{TT}$ by reverse-biasing a previously short-circuited p^+ -n junction and then illuminating the diode with photons of energy $h\nu$ such that $E_c - E_v > h\nu > E_c - E_T$, the resulting total change of the diode capacitance is obtained from (9.10) and (9.18) as

$$\Delta C = \text{const. } N_{TT}e_n^o/(e_p^o + e_n^o). \quad (9.19)$$

Hence, by measuring ΔC and the time constant $\tau = 1/(e_p^o + e_n^o)$ of the capacitance change for different photon energies, one can calculate the spectral distribution of the photoionisation cross-section $\sigma_n^o = e_n^o/\phi$ by using (9.19), since

$$\sigma_n^o = \Delta C \text{ (const. } N_{TT}\phi \tau)^{-1}. \quad (9.20)$$

Here, ϕ ($s^{-1} \text{ cm}^2$) is the photon flux used for the optical excitation. The measurements are particularly simple in the energy range $E_T - E_v > h\nu > E_c - E_T$, since for these energies $e_p^o = 0$ and absolute values of e_n^o are gained directly from the transient time constant τ .

The spectrum of e_p^o is obtained from similar measurements by choosing the initial conditions such that $n_T(0) = 0$. This can be achieved by first illuminating the sample with photons of energy $E_T - E_v > h\nu' > E_c - E_T$

prior to the measurements. If the junction is then illuminated with photons of energy $E_c - E_v > h\nu > E_T - E_v$, σ_p^o can be obtained from a similar relation to (9.20), including the same constant. Because, in the energy range $E_T - E_v > h\nu > E_c - E_T$, e_n^o is obtained directly from τ , the constant can be determined when ΔC is measured at the same photon energy and N_{TT} is known. It should be noted that if the defect is partially filled and then illuminated with photons of energy somewhat less than half the bandgap, the sign of ΔC gives information on the position of the defect, i.e. whether the defect is in the upper or lower half of the bandgap. This is an important advantage of the photocapacitance technique compared with photocurrent measurements.

Spectral distributions of photoionisation cross-sections can also be obtained from photocurrent methods. Considering that the steady-state photocurrent is given by

$$I_{ph} = qA(W - W_o)N_{TT}e_p^o e_n^o / (e_p^o + e_n^o), \quad (9.21)$$

where $W - W_o$ is the effective generation region [22], it is quite clear that, for a defect in the upper half of the bandgap, e_n^o can only be measured if $e_p^o < e_n^o$. On the other hand, there is no simpler and faster way of measuring optical emission rates with reasonable accuracy than by photocurrent measurements, since the optical emission rates are obtained directly and with high sensitivity. For the determination of e_p^o , other photocurrent techniques such as the dual-light-source steady-state photocurrent method [23] have to be used.

In addition to the already mentioned JSCTs for measuring optical properties, there are quite a number of steady-state measurement techniques. These techniques allow one to study defects at higher temperatures than can be used in transient measurements [23–25], for example, and/or to provide single exponential transients, which are not always obtained with conventional transient measurements owing to the free-carrier tail extending from the neutral region into the space charge region [26–28].

It is fair to say that spectral distributions of photoionisation cross-sections obtained with these techniques do not in general provide much information on the electronic structure of defects, because in most cases the spectra are rather smooth and the threshold energies are often broadened owing to competing processes. Other techniques have therefore been developed for deriving these defect properties from measurement methods with much higher resolution.

9.4 Optical Measurement Methods Other than Junction Space Charge Techniques

9.4.1 Photothermal Ionisation Spectroscopy

Because most of the defects in silicon are non-radiative, absorption and photothermal ionisation spectroscopy (PTIS), as well as admittance techniques,

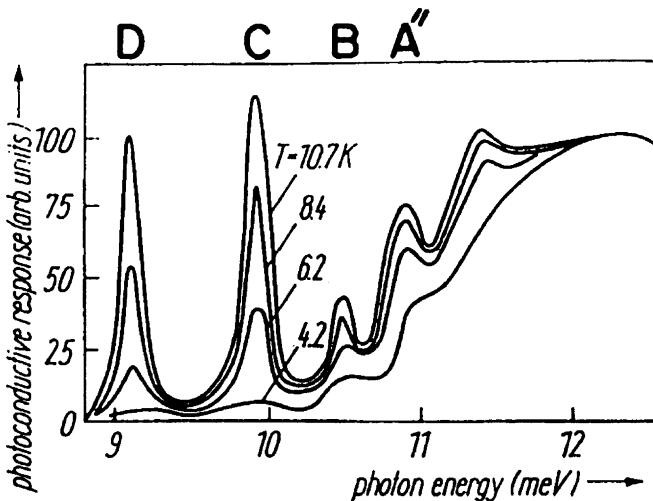


Fig. 9.3. PTIS spectra of indium-doped germanium (p-type) at different temperatures (from [29])

are of particular interest for optical studies of these defects. In contrast to the absorption measurements that are widely used, PTIS is less well known, though this technique has been used to study shallow centres with small binding energies for more than 30 years [29]. It took quite a while before this method was employed for studying so-called deep defects with larger binding energies. PTIS is a thermally enhanced photoconductivity method, based on the optical excitation of charge carriers from the ground state of a defect into excited states, followed by thermal ionisation of the excited defect. The absorption of a photon may or may not result in a positive PTIS signal, depending on whether or not the charge carrier is finally excited into the band continuum.

These basic properties of PTIS can be easily recognised in Fig. 9.3, which shows typical PTIS spectra of a shallow acceptor in germanium at different temperatures [29]. The defect (indium) has a binding energy of 11.7 meV. Transitions from the ground state into the valence-band continuum are seen as a broad, featureless spectrum at energies greater than the binding energy. On the low-energy side of the spectrum, five lines are observed. The signal of these lines increases with increasing temperature owing to two-step excitation via excited states, where the second excitation process is a thermal process. In contrast to absorption measurements, signals from shallower energy levels are therefore favoured over signals from deeper levels.

Another interesting difference between PTIS and absorption measurements is the high sensitivity of PTIS. Early investigations of shallow centres showed [29] that under certain conditions, especially in a particular tempera-

ture range, the PTIS signal is independent of the defect concentration down to very low values (10^5 cm^{-3} in germanium).

A further advantage of PTIS compared with absorption measurements is due to the fact that PTIS involves a two-step excitation process. This implies that PTIS allows one to study not only optical transitions, revealing the energy structure of the defect, but also the thermal properties of each excited state involved in the spectrum.

To illustrate the technique in more detail, the neutral substitutional selenium donor in silicon will be used as an example. The PTIS spectrum of the centre will be shown in Fig. 9.4. Since not all lines are observed in high resolution at one temperature, the spectrum between 2360 and 2500 cm^{-1} was recorded at 24.7 K and the spectrum between 2110 and 2230 cm^{-1} at 65.8 K. The spectrum consists of a continuum part at higher energies and several lines at lower energies that reflect the energy structure of the neutral selenium centre. Leaving out the s states, a comparison of the peak energies with the effective-mass approximation (EMA) shows (Fig. 9.5) that the structure of the spectrum is in perfect agreement with the EMA [1]. This implies, for example, that the peak at the second lowest energy corresponds to transitions from the ground state $1s(A_1)$ to the $2p_0$ state and that the peak with the highest energy is generated by transitions from the ground state to the $5p_{\pm}$ state. From a line spectrum such as the one shown in Fig. 9.4, the binding energy of the ground state is easily obtained, e.g. by measuring the excitation energy of the transition $1s(A_1) \rightarrow 2p_{\pm}$ and adding the EMA binding energy of the $2p_{\pm}$ state (6.40 meV), which gives 306.63 meV for the binding energy

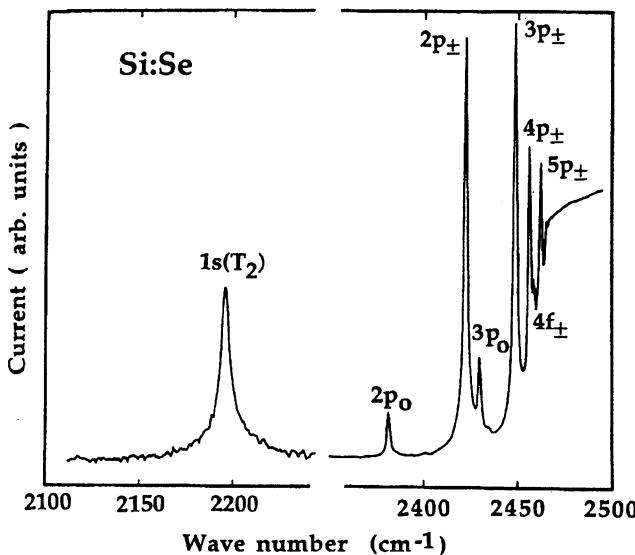


Fig. 9.4. PTIS spectra of Si:Se[°] (from [32])

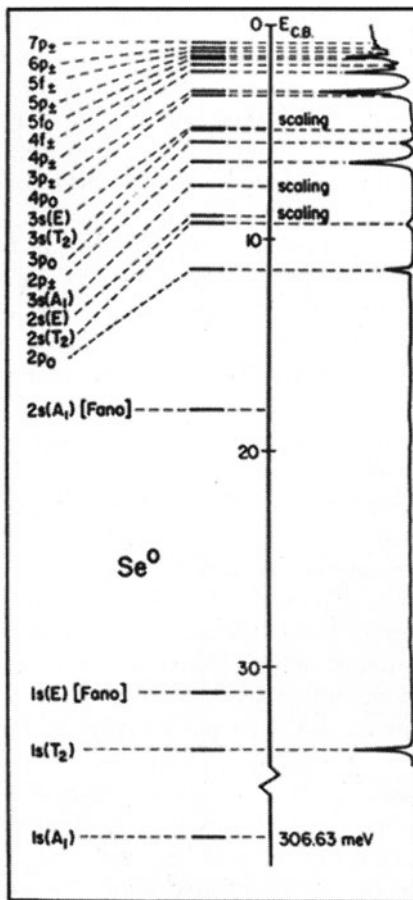


Fig. 9.5. Energy states of the first electron bound to a selenium donor in silicon. The spectrum shown can be obtained by PTIS or absorption

of the ground state. Once the ionisation limit is known, the binding energies of all excited states are readily derived.

Since the EMA binding energies of excited states increase with Z^2 , where Z is the core charge, the energy spacing of the excited states tells us whether the defect is neutral or charged, i.e. whether the defect is, for example, a single or a double donor. The line spectrum also reveals whether the optical transitions are donor- or acceptor-like (because they are different) and whether the defect is isolated or forms a complex. If, for example, the defect consists of a pair instead of an isolated substitutional chalcogen atom, the local symmetry is lowered from tetrahedral to trigonal, which causes a splitting of the s states, whereas the binding energies of the p, d and f states are more or less unaffected [30].

Neither these high-resolution optical measurements nor JSCTs give information on the chemical identity of the defect. This information is often obtained from ESR measurements in combination with other techniques, for example by correlating the ESR spectrum with the energy position of the defect when the sample is illuminated with monochromatic light of different energies. Owing to the different time dependences of the photon-induced valency change of the ESR active centre for different photon energies, the spectral distribution of the optical emission rate can be measured and be compared with the spectrum obtained from JSCTs [31].

Another technique for identifying defects is based on the fact that impurities with masses smaller than that of the host crystal atoms have, in general, vibrational frequencies well above the phonon frequency spectrum. This gives rise to vibrational modes with sharp spectral absorption bands in the infrared due to a strong spatial localisation. In conjunction with isotopic doping, it is then possible to identify the atoms involved in certain defects.

Moreover, using local-vibrational-mode (LVM) spectroscopy together with perturbations such as polarization of the probe light and uniaxial or hydrostatic stress allows one to determine the structure of defect complexes. It was the development of powerful Fourier transform spectrometers that made the fast expansion of LVM spectroscopy possible.

Examples of light impurities that have been successfully investigated in both elemental and compound semiconductors are hydrogen, carbon and oxygen. LVM spectroscopy is also used industrially for the control of various stages in the production of both silicon and integrated circuits.

More detailed treatments of LVM spectroscopy in semiconductors have been presented by Barker and Sievers [33], by Spitzer [34] and by Newman [35]. A further review has recently been published by Haller [36].

The thermal properties of the excited states are studied by measuring the PTIS spectra at different temperatures and by plotting the logarithm of the integrated signal of each peak versus $1/T$ (Fig. 9.6). From the slope of the straight line, the thermal activation energy ΔE^t (9.6) is obtained, provided (9.4) is valid [3]. If ΔH_n^c is similar for all excited states, a plot of ΔE^t versus the optical energy ΔE^0 needed for the excitation of an electron from the ground state into one of the excited states is expected to give a straight line with slope 1 (Fig. 9.7), intercepting the ΔE^0 axis at an energy given by the sum of ΔH_n^c and the binding energy of the ground state. If all data points lie on a straight line, this proves that all peaks of the PTIS spectrum originate from the same defect and that the capture process is similar for all excited states. Hence, if some of the data points do not lie on the straight line, this may imply that these states are correlated with a different defect or that the capture mechanism of these excited states differs from other excited states.

Two interesting conclusions can be drawn from the data presented in Fig. 9.6: (1) PTIS reveals thermal activation energies with much better ac-

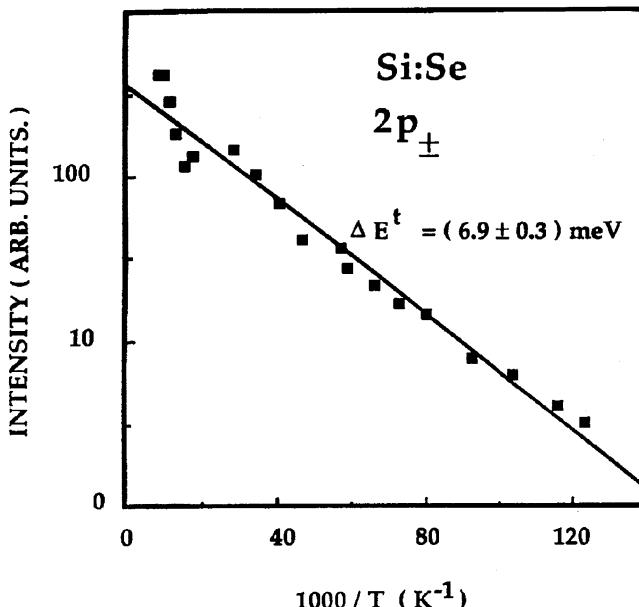


Fig. 9.6. Integrated amplitude of the $2p_{\pm}$ line for the neutral selenium donor in silicon as a function of reciprocal temperature (from [32])

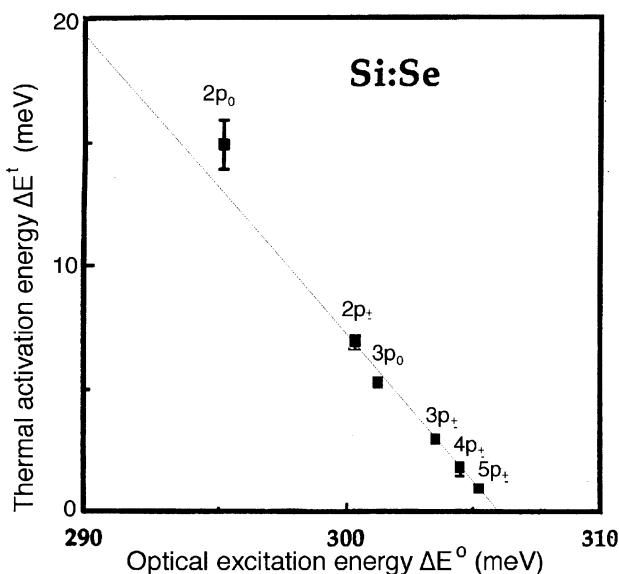


Fig. 9.7. Thermal activation energies of all lines higher than $1s(T_2)$ (see Fig. 9.4) as a function of optical excitation energy (from [32])

curacy than, for example, JSCTs. This is due to the fact that the PTIS technique inherently selects centres with very similar electronic properties. (2) The thermal activation energy ΔE^t obtained is very close to the optical binding energy, seemingly implying that the ΔH_n^c of higher excited states is small. This also explains why the regression line of Fig. 9.7 intercepts the ΔE^0 axis at an energy (306.5 meV) that is almost identical to the binding energy of the ground state. The fact that the experimental value for the $2p_0$ (and maybe also for the $1s(T_2)$) peak does not lie on this line suggests that different capture properties for these excited states may be valid.

PTIS spectra taken at different temperatures show [31] that the optical binding energies of all states, including the ground state, are almost temperature-independent and that any shift in the relative position of the peaks seems to be related to the temperature dependence of the dielectric constant.

9.4.2 Fourier Photoadmittance Spectroscopy

Fourier photoadmittance spectroscopy (FPAS) [37] combines the advantage of two well-known measurement techniques, namely admittance spectroscopy as applied to diodes [38] and Fourier spectroscopy as applied to photoconductors [39]. It has the signal-to-noise ratio and resolution of Fourier spectroscopy, and the sensitivity of junction techniques.

The spectrometer used in this measurement technique is a Fourier transform infrared (FTIR) spectrometer. The usual detector is replaced by the sample, which in this case is a diode. As the light intensity varies, an oscillating current is generated in the external circuit of the diode owing to excitations from and capture into localised energy states within the forbidden energy gap. Using a diode as a detector in an FTIR spectrometer implies that the detector signal is a combination of a large number of oscillating currents, each current component having a frequency proportional to the energy of the corresponding component of the incident light spectrum. This signal is stored in the computer of the spectrometer, and after a suitable period of data collection is Fourier transformed into the spectral response of the diode to the known incident spectrum. It has been shown that the alternating current in the external circuit of the diode is proportional to the optical emission rate $e_{n,p}^o$ if transitions to only one of the energy bands are involved [40]. The experimental conditions can often be arranged such that deviations from linearity between the response and the emission rates are small.

FPAS has several advantages compared with conventional absorption and junction techniques. Since the light passes through only a very short distance before it becomes active, its intensity is much less affected by phonon absorption than in absorption measurements. Compared with conventional optical junction techniques, it is much faster, and the resolution is considerably better. It is the only junction technique that allows the optical study of energy

levels smaller than 0.2 eV. Furthermore, measurements on diodes involve depletion regions or their edge zones. The small width of the active part of the depletion region makes space charge regions particularly suitable for the study of defects in thin layers deposited epitaxially on a thicker substrate or of defects introduced by shallow diffusion. In this case, suitable junctions are obtained either by depositing a metal film on the layer that is to be studied, or by doping a thin surface layer p⁺ or n⁺.

When very high resolution is not required, the FTIR spectrometer can be replaced by any other spectrometer with proper modulation of the monochromatic light. The measurement technique is then called photoadmittance spectroscopy (PAS) [41].

References

1. W. Kohn: Solid State Phys. **5**, 257 (1957)
2. W.E. Krag, H.J. Zeiger: Phys. Rev. Lett. **8**, 485 (1962)
3. G. Björklund, H.G. Grimmeiss: Physica Status Solidi **42**, K1 (1970)
4. C.T. Sah, L. Forbes, L.L. Rosier and A.F. Tasch Jr.: Solid State Electron. **13**, 759 (1970)
5. D.V. Lang: J. Appl. Phys. **45**, 3023 (1974)
6. G.W. Ludwig, H.H. Woodbury: *Solid State Physics*, vol. 13, ed. by F. Seitz, D. Turnbull (Academic Press, New York 1962) p. 223
7. C.A.J. Ammerlaan, P.T. Huy: Solid State Phenomena **85-86**, 353 (2002)
8. A. Thilderkvist, M. Kleverman, G. Grossmann, H.G. Grimmeiss: Phys. Rev. B **49**, 14270 (1994)
9. K. Bergman, G. Grossmann, H.G. Grimmeiss, M. Stavola, R.E. McMurray Jr.: Phys. Rev. B **39**, 1104 (1989)
10. O. Engström, A. Alm: Solid State Electron. **21**, 1571 (1978)
11. M. Lax: Phys. Rev. **119**, 1502 (1960)
12. V.N. Abakumov, V.I. Perel, I.N. Yassievich: Sov. Phys. Semicond. **12**, 1 (1978)
13. R.M. Gibb, G.J. Rees, B.W. Thomas, B.L.H. Wilson, B. Hamilton, D.R. Wight, N.F. Mott: Philos. Mag. **36**, 1021 (1977)
14. J.A. Pals: Solid State Electron. **17**, 1139 (1974)
15. R. Williams: J. Appl. Phys. **37**, 3411 (1966)
16. H.G. Grimmeiss, C. Ovrén: J. Phys. E **14**, 1032 (1981)
17. G.L. Miller, D.V. Lang, L.C. Kimmerling: Annu. Rev. Mater. Sci. (1977) p. 377
18. L.C. Kimmerling, J.L. Benton: Appl. Phys. Lett. **39**, 410 (1981)
19. K. Hofmann, M. Schulz: Appl. Phys. A **33**, 19 (1984)
20. H. Pettersson, H.G. Grimmeiss: Phys. Rev. B **42**, 1381 (1990)
21. L. Montelius, H.G. Grimmeiss: Semicond. Sci. Technol. **3**, 847 (1988)
22. S. Braun, H.G. Grimmeiss: J. Appl. Phys. **44**, 2789 (1973)
23. S. Braun, H.G. Grimmeiss: J. Appl. Phys. **45**, 2658 (1974)
24. G. Björklund, H.G. Grimmeiss: Solid State Electron. **14**, 589 (1971)
25. H.G. Grimmeiss, N. Kullendorff: J. Appl. Phys. **51**, 5852 (1980)
26. A. Zylbersztejn: Appl. Phys. Lett. **33**, 200 (1978)
27. H.G. Grimmeiss, E. Janzén, B. Skarstam: J. Appl. Phys. **51**, 4212 (1980)

28. H.G. Grimmeiss, E. Janzén, B. Skarstam, A. Lodding: *J. Appl. Phys.* **51**, 6238 (1980)
29. T.M. Lifshits, N.P. Likhtman, V.I. Sidorov: *Fiz. Tekh. Poluprovodn.* **2**, 782 (1968)
30. E. Janzén, R. Stedman, G. Grossmann, H.G. Grimmeiss: *Phys. Rev. B* **29**, 1907 (1984)
31. H.G. Grimmeiss, E. Janzén, H. Ennen, O. Schirmer, J. Schneider, R. Wörner, C. Holm, E. Sirtl, P. Wagner: *Phys. Rev. B* **24**, 4571 (1981)
32. J.-O. Fornell: Thesis, University of Lund (1989)
33. A.S. Barker Jr., A.J. Sievers: *Rev. Mod. Phys.* **47**, Suppl. 2 (1975)
34. L.H. Solnik, W.G. Spitzer, A. Kahan, R.G. Hunsperg: *J. Appl. Phys.* **42** (13), 5223 (1971)
35. R.C. Newman: *Infrared Studies of Crystal Defects* (Taylor & Francis, London 1973)
36. E.E. Haller: *Mater. Res. Soc. Symp. Proc.* **378**, 547 (1995)
37. E. Janzén, K. Larsson, R. Stedman, H.G. Grimmeiss: *J. Appl. Phys.* **53**, 7520 (1982)
38. D.L. Losee: *Appl. Phys. Lett.* **21**, 54 (1972)
39. M.S. Skolnick, L. Eaves, R.A. Stredling, J.C. Portal, S. Askenazy: *Solid State Commun.* **15**, 1403 (1974)
40. K. Larsson: *J. Phys E Sci. Instrum.* **20**, 1480 (1987)
41. M. Kleverman, E. Janzén, H.G. Grimmeiss: *Solid State Commun.* **46**, 895 (1983)
42. *Landolt-Börnstein*, New Series vol. 22, ed. by O. Madelung, M. Schulz (Springer-Verlag, Berlin, Heidelberg 1989) p. 68

10 Silicon and Its Vital Role in The Evolution of Scanning Probe Microscopy

F.J. Giessibl

10.1 Introduction

Silicon has played an important role in the development of scanning probe microscopes. The 7×7 reconstruction of silicon has been a challenge for surface scientists for more than two decades. In 1983, the scanning tunneling microscope (STM), invented by Gerd Binnig and Heinrich Rohrer, yielded the first atomically resolved image of Si (111) (7×7) in real space. The information contained in this data has helped to develop the dimer–adatom–stacking-fault model of Takayanagi et al. The Si (111) (7×7) surface has been considered a touchstone for a different scanning probe microscope: the atomic force microscope (AFM). While STM images of the Si (111) (7×7) surface were achieved a year after the introduction of the STM, the AFM took almost a decade to meet the challenge of this excitingly complicated surface. While the initial AFM data did not give new information about the silicon surface, the quest for imaging this complicated surface has served as a driving force for perfecting the AFM. The resolution of the AFM has been increased and the Si (111) (7×7) surface is now a standard test sample for the AFM. Because the surface is known well, it can be used to learn more about the three-dimensional structure of the probe tip of the microscope. Silicon is both an exciting object of study and an important material for scanning probe microscopes.

10.2 Silicon as a Benchmark for Scanning Probe Microscopes

In 1959, Schlier and Farnsworth reported their low-energy electron diffraction (LEED) experiments on the surface of Si (111) [1]. After heating the surface to 900°C , they discovered that the surface displays additional scattering peaks in the LEED pattern. These additional peaks are caused by a reconstruction of the Si surface, where the new unit cell is 7×7 times as large as the bulk terminated structure. Because of the large size of the unit cell, the structure of this surface stood as a tremendous challenge for surface scientists for more than two decades. In 1982, Binnig, Rohrer et al. [2] introduced the scanning tunneling microscope (STM). Figure 10.1 shows the principle of

this remarkable instrument. A sharp tip is mounted on a piezoelectric tripod scanner. This scanner allows one to move the tip in three-dimensional space with atomic precision. A coarse positioning device (not shown in Fig. 10.1) moves the sample within reach of the tripod scanner. When the tip and sample (with a relative bias voltage V_t) approach within a distance of a few atomic diameters, a tunneling current I_t flows. The tunneling current increases roughly exponentially with distance, by a factor of 10 for every 100 picometers of distance reduction. This rapid, monotonic distance dependence allows a simple feedback arrangement which adjusts the z position in order to keep the current constant. Scanning the tip in the x - y plane and recording $z(x, y, I_t = \text{const.})$ yields a topographic image of the surface. Because of the rapid decay of the tunneling current with distance, even relatively blunt tips yield atomic resolution easily. Only one year after the introduction of the STM, Binnig et al. determined the positions of the surface atoms of this iconic surface (Fig. 10.2) with a scanning tunneling microscope [3]. This fantastic result immediately demonstrated the utility of the STM as a tool for surface science, and in 1986 Binnig and Rohrer received the Nobel Prize in physics for the invention of the STM.

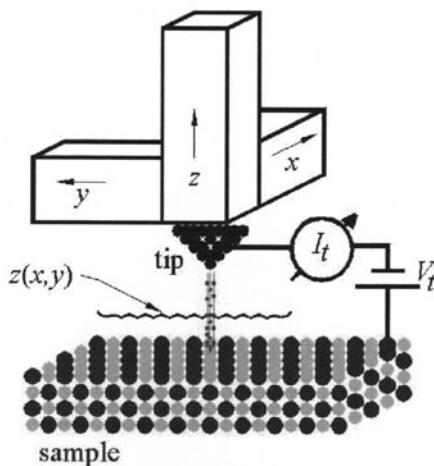


Fig. 10.1. The scanning tunneling microscope (STM). The STM consists of a piezoelectric scanner which allows one to move a sharp tip in three dimensions and scan a surface. If the tip is close enough to a sample biased at a voltage V_t , a tunneling current I_t (little dots) can flow between sample and tip. This current increases rapidly with decreasing tip-sample distance, yielding a superb input signal for an electronic feedback control that keeps the tip-sample distance constant. Recording the position of the tip $z(x, y, I_t = \text{const.})$ as it is scanned in the x - y plane yields a topographic map of the surface that, ideally, resolves every single sample atom

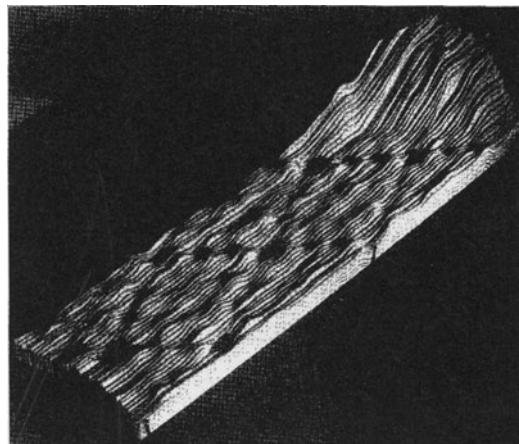


Fig. 10.2. First atomic image of the Silicon 7×7 reconstruction in real space by Binnig et al. (1983). The diamond-shaped unit cell, with 12 adatoms and deep corner holes, is shown. In 1983, computer graphics were not easily available, and the image was created by cutting the traces of a chart recorder from cardboard and arranging the pieces to build a 3D model of the surface

Silicon condenses in the diamond structure with a cubic lattice constant of $a_0 = 0.543 \text{ nm}$ at $T = 0 \text{ K}$ and 0.54355 nm at $T = 300 \text{ K}$ [4]. The unit vectors of the bulk terminated Si (111) surface have a length of $v = a_0/\sqrt{2} = 0.384 \text{ nm}$ and an angle of 60° ; in the reconstructed surface, the unit vectors have a length of $w = 7 \times 0.384 \text{ nm} = 2.688 \text{ nm}$ (see Fig. 10.3). The reconstruction affects not only the surface atomic layer, but also the topmost four layers, and the new unit cell contains approximately 200 atoms. Because of this large size of the unit cell, the determination of the atom positions was a tremendous scientific challenge that remained unsolved for more than two decades. In 1985, the now commonly adopted dimer–adatom–stacking-fault (DAS) model was finally suggested by Takayanagi et al. [5]. According to this model (see Fig. 10.3), six adatoms are situated in each half of the unit cell. The adatoms are bound by covalent bonds formed by the overlap of sp^3 hybrid orbitals. In the bulk, the hybrid orbitals of neighboring atoms overlap and form an electron pair. At the surface, one of the four sp^3 hybrid orbitals is pointing perpendicular to the surface and forms a dangling bond.

Imaging Si (111) (7×7) was considered a standard test for the resolution capability and surface-science compatibility of an STM. Other orientations of the Si surface, such as Si (001) (2×1) surfaces, were studied later, and spectroscopy methods with $I_t(V_t)$ or $I_t(z)$ measurements were performed; see [6, 7] and references therein.

The STM needs the flow of a tunneling current as a control signal for the operation of the microscope. In 1985, Binnig invented the atomic force microscope. In the AFM, the forces that act between the tip and the sample replace

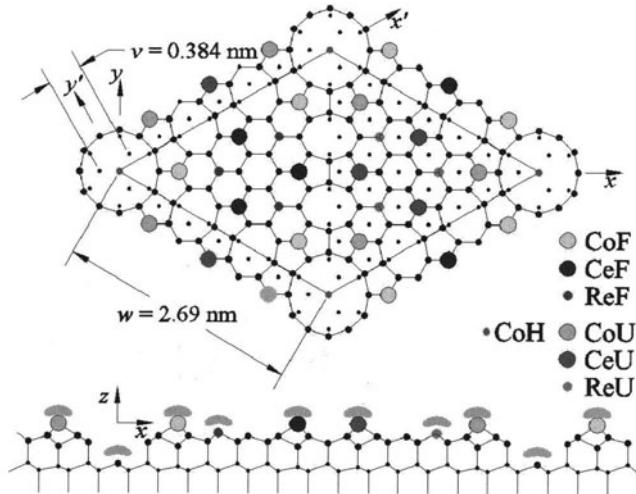


Fig. 10.3. Top view and cross section ($y = 0$) of the dimer–adatom–stacking-fault (DAS) model of Si (111) (7×7). The unreconstructed surface lattice has a lattice constant of 384 pm. By forming the 7×7 reconstruction, most of the dangling bonds of the surface atoms are saturated by an adatom such that the total number of dangling bonds per unit cell is reduced from 49 to 19. Twelve adatoms and a corner hole per unit cell are the striking features of this reconstruction. The adatoms fall into four symmetry classes: corner faulted (CoF), center faulted (CeF), corner unfaulted (CoU), and center unfaulted (CeU). The 19 remaining dangling bonds originate from the 12 adatoms, the 6 remaining atoms (ReF, ReU), and the atom in the center of the corner hole (CoH)

the tunneling current as the control signal. In 1986, Binnig, Quate, and Gerber introduced the first realization of the AFM principle and imaged steps on an Al_2O_3 surface [8]. Figure 10.4 shows the principle of the AFM. The tunneling tip is replaced by a tip mounted on a cantilever. Forces between the tip and the sample cause bending of the cantilever, allowing one to measure the forces.

Obtaining atomic resolution with an STM is relatively simple because of the physical properties of the tunneling current: short range and monotony. Figure 10.5 shows a typical plot of the distance dependences of the tunneling current (4), short-range force (1), and long-range force (2).

Because of the monotonic distance dependence of the tunneling current, the implementation of a feedback control circuit is simple in the STM. The tunneling current is simply fed into a logarithmic amplifier. A control circuit compares the logarithm of the current with the logarithm of the current set-point. If the actual current is larger than the setpoint, the tip is withdrawn; if the current is too small, the tip is approached further. Because the tip–sample force F_{ts} is not monotonic, this simple principle does not work in the AFM.

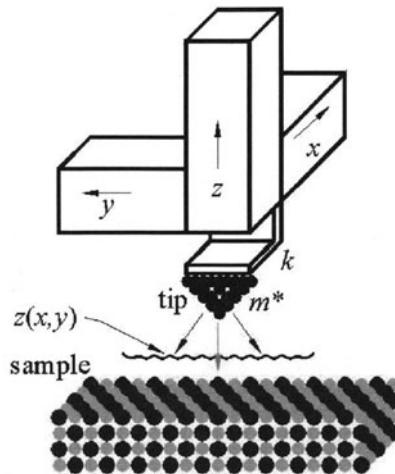


Fig. 10.4. Atomic force microscope (AFM). The operating principles of the AFM and STM are similar. In the AFM, the tip is mounted on a cantilever with stiffness k and effective mass m^* . Tip-sample forces F_{ts} cause a cantilever deflection $q' = F_{ts}/k$. In contrast to the tunneling current in the STM, F_{ts} is neither purely short-range nor monotonic, but may have attractive and repulsive components. Chemical forces have a short range and act mainly between the tip's front atom and the sample atom next to it (gray arrow), but long-range forces, caused for example by van der Waals- or electrostatic interaction (dark arrows), are also typically present. Under many experimental conditions, the long-range forces dominate over the short-range forces in magnitude

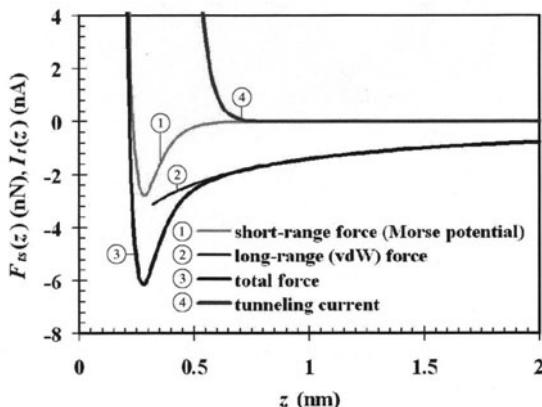


Fig. 10.5. Tunneling current, and short-range, long-range, and total tip-sample force (typical) as a function of tip-sample distance. The tunneling current has a short range and a monotonic distance dependence, which simplifies feedback. The tip-sample force is composed of short- and long-range components and is not monotonic

Imaging the Si (111) (7×7) surface with atomic resolution was instrumental in the success of the STM, and imaging this surface by AFM was a touchstone for the force microscope. However, with reactive surfaces such as silicon, more problems in addition to the long-range and nonmonotonic properties of the tip–sample forces illustrated in Fig. 10.5 arose. The chemical bonds which emerged between the AFM tip and the Si surface were strong, and scanning created heavy wear and tear on both tips and samples. Howald et al. [9] covered the tip of the cantilever with polytetrafluoroethylene and managed to reduce the sticking forces such that the 7×7 unit cell could be imaged. Individual atoms, however, were not resolved in these experiments. Figure 10.6 shows the first result of atomic resolution on the Si surface [10], obtained in 1994 with a dynamic AFM mode explained below. Similarly to the first STM results, only a small area was imaged with atomic resolution, and the individual atoms appeared as noisy bumps (see inset in Fig. 10.6).

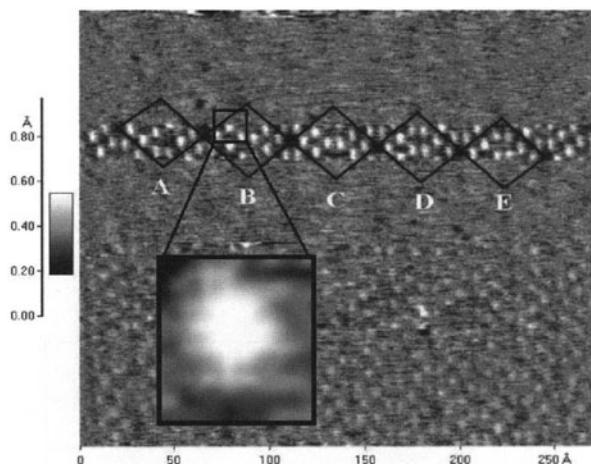


Fig. 10.6. First AFM image of Si (111) (7×7). Image size 27 nm × 19.5 nm. A cantilever with stiffness $k = 17 \text{ N/m}$, eigenfrequency $f_0 = 114\,224 \text{ Hz}$, and quality factor $Q = 28\,000$ was oscillated at an amplitude $A = 34 \text{ nm}$. The electrostatic bias between the tip and the sample was zero in order to minimize long-range electrostatic forces. The attractive forces led to a frequency change of $\Delta f = -70 \text{ Hz}$ and the image was recorded at this constant negative frequency shift. The black diamonds A–E indicate the five unit cells that are clearly resolved in this image. The variation in the image quality is due to a change in the tip – the front atom of the tip has changed its configuration during the scan. The inset shows a magnified view of a single adatom

The experimental technique that was used in Fig. 10.6 was introduced by Albrecht et al. in 1991 [11] and initially used for magnetic force microscopy. This technique, “frequency-modulation atomic force microscopy”,

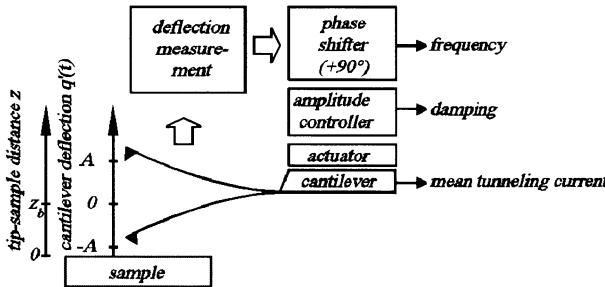


Fig. 10.7. Principle of frequency-modulation AFM. A cantilever (shown at its upper and lower turnaround points) is driven at a constant amplitude A by applying positive feedback. The frequency adjusts to the value given in (10.3). The energy required to maintain a constant amplitude is a measure of the hysteresis in the tip sample forces. For conducting samples, the tunneling current is a third physical observable accessible by FM-AFM

is explained in Fig. 10.7. Rather than the dc deflection of the cantilever being measured, the cantilever is subjected to positive feedback such that it oscillates at its eigenfrequency f_0 with an adjustable but constant amplitude A . The eigenfrequency of a harmonic oscillator is given by

$$f_0 = \frac{1}{2\pi} \sqrt{\frac{k^*}{m^*}}, \quad (10.1)$$

where k^* is its effective spring constant and m^* its effective mass. Tip-sample forces change the effective spring constant, and the stiffness of the tip-sample bond k_{ts} has to be added to the spring constant of the cantilever, so that $k^* = k + k_{ts}$. If $k_{ts} \ll k$, the square root in (10.1) can be expanded as a Taylor series and the linear term in the expansion yields the dominant correction to (10.1). The frequency then changes by Δf , which is given by

$$\Delta f = \frac{f_0}{2k} k_{ts} \quad (10.2)$$

if the tip-sample force gradient is constant during the oscillation cycle. However, in experiments involving amplitudes of some tens of nm, k_{ts} varies by orders of magnitude over the z interval $[z - A, z + A]$, and (10.2) has to be generalized to

$$\Delta f(z) = \frac{f_0}{k\pi A^2} \int_{-A}^A k_{ts}(z + \zeta) \sqrt{A^2 - \zeta^2} d\zeta, \quad (10.3)$$

as shown in [12] and references therein.

The frequency modulation technique, with parameters similar to the ones used in Fig. 10.6, was successful in imaging Si, other semiconductors, metals, and insulators [13–18].

10.3 Silicon as a Material for AFM Cantilevers

The cantilever is the central element of an AFM and the part that distinguishes it from its predecessor, the STM. The first cantilevers were built from gold foils [8] with attached diamond tips. Later, micromachined silicon cantilevers became popular. The first micromachined cantilevers used a V-shaped amorphous SiO_2 film etched on a Si wafer as the cantilever. Later, cantilevers made from Si_3N_4 films on Si wafers became popular because Si_3N_4 is more durable than SiO_2 . Today, most cantilevers are made from single-crystal silicon with integrated tips.

Figure 10.8 shows a few examples of cantilevers. Figure 10.8a shows a silicon cantilever (Nanosensors GmbH), Fig. 10.8b depicts a cantilever (Olympus) where the tip can be viewed from above with an optical microscope, Fig. 10.8c shows a self-sensing piezoresistive cantilever made from single-crystal silicon, and Fig. 10.8d displays a “cantilever” made from a crystalline SiO_2 (quartz) tuning fork.

In the first AFM [8], the deflection of the cantilever was measured with a tunneling tip – the back of the cantilever was used as an electrode, and a sharp metal tip was brought within tunneling distance to it. Later AFMs used optical beam-deflection or interferometric methods [24]. In doped silicon, the piezoresistive effect is quite strong, which allows one to build self-sensing cantilevers [21] as shown in Fig. 10.8c. In crystalline SiO_2 , the piezoelectric effect also allows one to make self-sensing cantilevers [22, 23] (Fig. 10.8d).

10.4 The $\text{Si}(111)$ (7×7) Surface as a Probe for STM and AFM Tips

In the STM and AFM, the atomic images are convolutions of the electronic states of the tip and the sample [7]. Thus, the atomic state of the tip needs to be known in order to fully characterize the image. Fortunately, in most cases it is not necessary to know exactly the electronic state of the tip. In typical experiments, shaping the tip is done *in situ*, for example by performing mild collisions with the sample, applying voltage pulses, etc., until “good” images result. Because of the topographical features of the Si (111) (7×7) surface – deep corner holes, and widely spaced adatoms with a precisely known electronic configuration – Si is an excellent test surface to study the electronic structure of the microscope tip.

After a systematic study of the factors minimizing noise in the AFM, a cantilever with nearly ideal physical properties was created (“qPlus sensor”, see Fig. 10.8d). In 2000, this cantilever was used to image Si (111) (7×7). Because the electronic configuration of the adatoms comprising the surface atom layer is 3sp^3 oriented perpendicular to the surface (see Fig. 10.3), adatom images that are rotationally symmetric with respect to the vertical axis were expected. Instead, the images of individual atoms (Fig. 10.9) show a peak

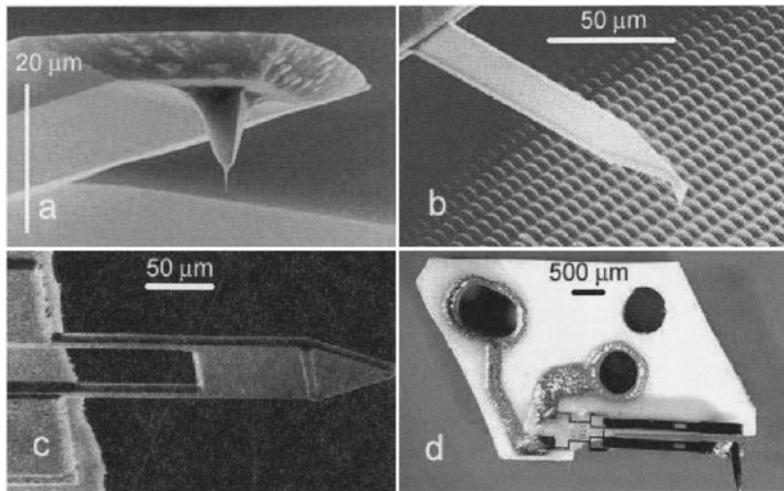


Fig. 10.8. Four types of force sensors (cantilevers). Forces on the tip cause a deflection of the cantilever. The deflection is measured optically with a beam-deflection or interferometric method in (a) and (b) and electrically in (c) and (d). (a) Micromachined silicon cantilever with an integrated tip pointing in the [001] direction [19]. (b) Another micromachined silicon cantilever with an integrated tip pointing in the [001] direction [20]. The tip is at the very end of the cantilever, such that the sample area next to the tip can be viewed in an optical microscope. (c) Self-sensing piezoresistive silicon cantilever with an integrated tip pointing in the [001] direction. The cantilever changes its resistance when it is deflected. Making the piezoelectric cantilever part of a Wheatstone bridge yields a simple way to create an electrical force signal [21]. (d) Cantilever based on a SiO_2 (quartz) tuning fork (“qPlus sensor”) [22, 23]. Owing to the piezoelectric effect in SiO_2 , a deflection of the sensor causes charges at the surfaces of the deflected prongs. These charges are collected by electrodes and measured with a current-to-voltage converter

consisting of two subpeaks [25]. Because of the shape and the spacing of the two subpeaks, an ordinary double-tip effect has to be ruled out. The AFM image is created by a convolution of tip and sample states, and the sample state is known to be an sp^3 orbital pointing in the z direction as shown in Fig. 10.10. Figure 10.9 is thus explained as an image of the tip orbitals. More recent work has also used the Si surface to probe the electronic structure of metal atoms with d and f valence shells [26].

Figures 10.9 and 10.10 emphasize the importance of the tip structure. Marcus et al. [27] have created ultrasharp Si tips oriented in the [001] direction and imaged them by transmission electron microscopy (Fig. 10.11). Interestingly, the crystal lattice appears to continue up to the very end of the tip. While Si [001] tips can be etched to form extremely sharp tips, the choice of the [001] crystal direction may not be optimal. Figures 10.12a,b show bulk-like terminated Si tips which point in the [001] and the [111] direction, respectively.

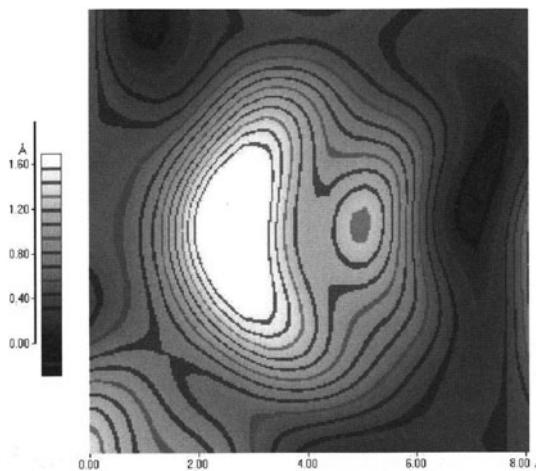


Fig. 10.9. AFM image of a single adatom on Si (111) (7×7) reflecting the electronic structure of the tip atom

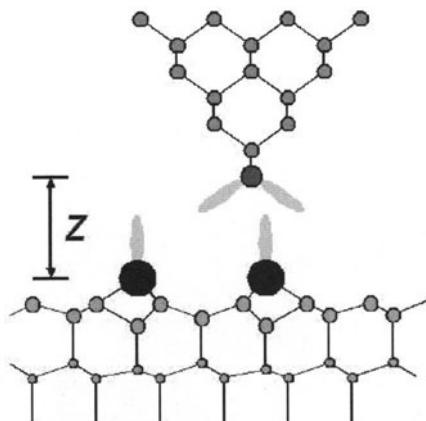


Fig. 10.10. Ball-and-stick model of atomic and orbital arrangements to explain Fig. 10.9. Two peaks per sample atom are expected to occur in this situation because Si strongly favors a tetrahedral bonding symmetry

The manufacturing process for Si cantilevers is easier for integrated tips pointing in the [001] direction than for tips pointing in the [111] direction. However, tips oriented in the [111] direction should be more stable, because the front atom has three bonds to the shaft of the tip and only one dangling bond. Moreover, it should be easy to make sharp [111] tips, because the natural cleavage planes in Si are $\{111\}$ and three planes that are not parallel always meet in a single point, i.e. an atomically sharp tip is expected for

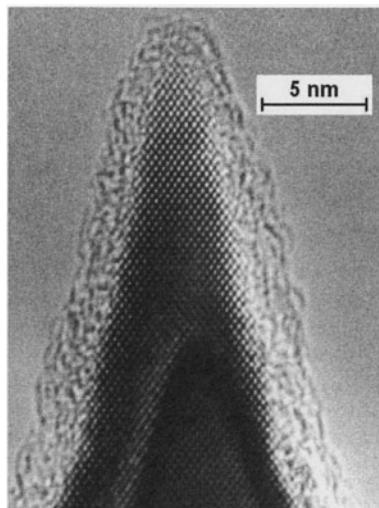


Fig. 10.11. Transmission electron micrograph of an extremely sharp silicon tip oriented in the [001] direction. The native oxide has been etched away with hydrofluoric acid before imaging. The 1.5–2.0 nm thick coating on the tip is mostly due to hydrocarbons which have been polymerized by the electron beam. Interestingly, the crystal structure appears to remain bulk-like up to the apex of the tip

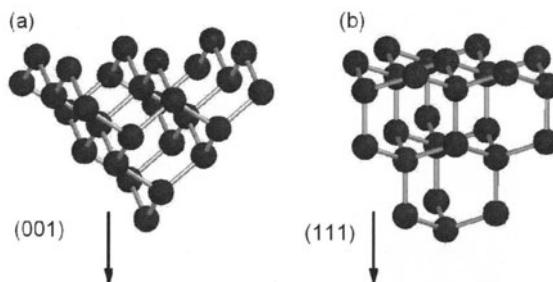


Fig. 10.12. Models of atomic arrangements for bulk-like terminated silicon tips, pointing in the (001) direction (a) and in the (111) direction (b)

tips limited by $\{111\}$ planes. The orientation and coordination of the front atom of a [111] tip can even be retained when the sidewalls reconstruct to Si $\{111\}$ (7×7) planes. Figure 10.13 shows an example of such a tip. This tip is cleaved from a single-crystal Si wafer and glued onto a qPlus sensor (see Fig. 10.8d). The front atom of such a tip should be bonded with great strength, and images of the Si(111) (7×7) surface obtained with such a tip should show only a single maximum per adatom. Figure 10.14 was acquired with an AFM using a tip of the kind shown in Fig. 10.13. As expected, every adatom shows a single peak. The contour lines in the adatom images show

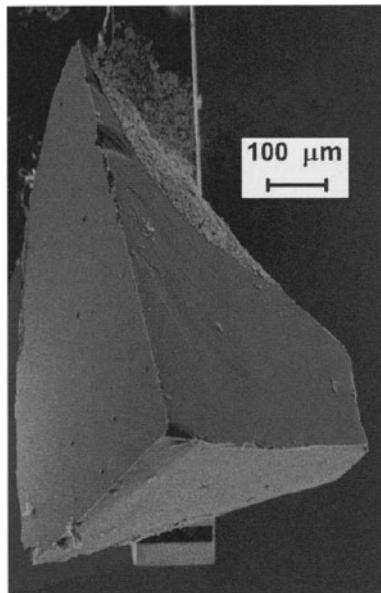


Fig. 10.13. Scanning electron micrograph of a cleaved single-crystal silicon tip attached to the free prong, of a qPlus sensor (see Fig. 10.8d). The rectangular section is the end of the free prong with a width of $130\text{ }\mu\text{m}$ and a height of $214\text{ }\mu\text{m}$. The tip points in the $[111]$ direction and is bounded by $(\overline{1}\overline{1}1)$, $(1\overline{1}\overline{1})$, and $(\overline{1}1\overline{1})$ planes

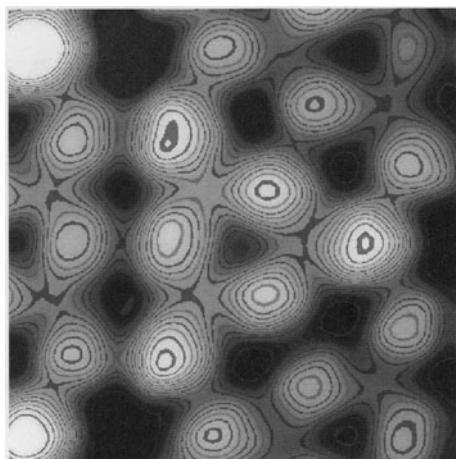


Fig. 10.14. Image of a $\text{Si}(111)$ (7×7) surface imaged with a qPlus sensor. Parameters: $k = 1800\text{ N/m}$, $A = 0.25\text{ nm}$, $f_0 = 14\,772\text{ Hz}$, $\Delta f = +4\text{ Hz}$. Image size: 4 nm lateral, 140 pm vertical [28]

that the tip was not aligned perfectly perpendicular to the surface. There are still angularly dependent force contributions, as indicated by the nonspherical contour profiles. The remaining atoms also show up in this image: between corner adatoms and center adatoms, the surface appears much shallower than between the three center adatoms, where there is no other atom in between (the distances between neighboring center adatoms are equal to the distance between corner and center adatoms). Figure 10.14 was even recorded with positive frequency shifts, i.e. repulsive tip-sample forces.

In summary, a tight link between silicon and scanning probe microscopes has been identified. The silicon surface started as a challenge for the STM and AFM, and later became a launchpad for the success of the STM and AFM. Today, it is a perfect test sample to study the electronic states of the tip. But silicon has more roles in probe microscopies: it serves as a material for cantilevers, and of course the electronics necessary to operate scanning probe microscopes could not do without silicon.

Acknowledgments

The author thanks G. Binnig, R.B. Marcus, O. Ohlsson (Nanosensors), and T. Akitoshi (Olympus) for permission to use figures, and J. Mannhart for support. This work was funded by the BMBF (project no. EKM 13N6918).

References

1. R.E. Schlier, H.E. Farnsworth: *J. Chem. Phys.* **30**, 917 (1959)
2. G. Binnig, H. Rohrer, C. Gerber, E. Weibel: *Phys. Rev. Lett.* **49**, 57 (1983)
3. G. Binnig, H. Rohrer, C. Gerber, E. Weibel: *Phys. Rev. Lett.* **50**, 120 (1983)
4. *Landolt-Börnstein, Numerical Data and Functional Relationships in Science and Technology*, Vol. 17a, ed. by O. Madelung, M. Schultz, H. Weiss (Springer, Berlin 1982) p. 370
5. K. Takayanagi, Y. Tanishiro, S. Takahashi: *J. Vac. Sci. Technol. A* **3**, 1502 (1985)
6. R. Becker, R. Wolkow. In: J.A. Stroscio, W.J. Kaiser (eds.): *Scanning Tunneling Microscopy* (Academic Press, San Diego 1993) pp. 149–224
7. C. J. Chen: *Introduction to Scanning Tunneling Microscopy* (Oxford University Press, New York 1993)
8. G. Binnig, C.F. Quate, C. Gerber: *Phys. Rev. Lett.* **56**, 930 (1986)
9. L. Howald, R. Lüthi, E. Meyer, H.-J. Güntherodt: *Phys. Rev. B* **51**, 5484 (1995)
10. F.J. Giessibl: *Science* **267**, 68 (1995)
11. T.R. Albrecht, P. Grütter, D. Horne, D. Rugar: *J. Appl. Phys.* **69**, 668 (1991)
12. F.J. Giessibl: *Appl. Phys. Lett.* **78**, 123 (2001)
13. S. Kitamura, M. Iwatsuki: *Jpn. J. Appl. Phys.* **34**, L145 (1995)
14. *Proceedings of the First International Workshop on Non-contact Atomic Force Microscopy* (Osaka, July 21–23, 1998), ed. by S. Morita and M. Tsukada, *Appl. Surf. Sci.* **140**, 243 (1999)

15. *Proceedings of the Second International Workshop on Non-contact Atomic Force Microscopy* (Pontresina, September 1–4, 1999), ed. by R. Bennewitz, C. Gerber, E. Meyer, *Appl. Surf. Sci.* **157**, 207 (2000)
16. *Proceedings of the Third International Conference on Non-contact Atomic Force Microscopy* (Hamburg, July 16–19, 2000), ed. by U.D. Schwarz, H. Hölscher, R. Wiesendanger, *Appl. Phys. A* **72**, S1 (2001)
17. *Proceedings of the Fourth International Conference on Non-contact Atomic Force Microscopy* (Kyoto, September 1–5, 2001), ed. by M. Tsukada, S. Morita, *Appl. Surf. Sci.* **188**, 231 (2002)
18. S. Morita, R. Wiesendanger, E. Meyer (eds.): *Noncontact Atomic Force Microscopy* (Springer, Heidelberg 2002)
19. O. Wolter, T. Bayer, J. Greschner: *J. Vac. Sci. Technol.* **9**, 1353 (1991)
20. Olympus Optical Co. Ltd., Tokyo, Japan
21. M. Tortonese, R.C. Barrett, C.F. Quate: *Appl. Phys. Lett.* **62**, 834 (1993)
22. F.J. Giessibl: *Appl. Phys. Lett.* **73**, 3956 (1998)
23. F.J. Giessibl: *Appl. Phys. Lett.* **76**, 1470 (2000)
24. D. Sarid: *Scanning Force Microscopy* (Oxford University Press, New York 1994)
25. F.J. Giessibl, S. Hembacher, H. Bielefeldt, J. Mannhart: *Science* **289**, 422 (2000)
26. M. Herz, F.J. Giessibl, J. Mannhart: unpublished
27. R.B. Marcus, T.S. Ravi, T. Gmitter, K. Chin, D. Liu, W.J. Orvis, D.R. Ciarlo, C.E. Hunt, J. Trujillo: *Appl. Phys. Lett.* **56**, 236 (1990)
28. F.J. Giessibl, H. Bielefeldt, S. Hembacher, J. Mannhart: *Ann. Phys. (Leipzig)* **10**, 887 (2001)

Part V

Doping Silicon

11 Defects, Diffusion, Ion Implantation, Recrystallization, and Dielectrics

E.F. Krimmel

11.1 Introduction

Ideally pure single-crystal silicon shows an intrinsic electrical conductivity, which is low at room temperature and rises with increasing temperature. To produce devices, well-defined parts of the silicon specimen, usually in the form of a wafer, must exhibit a well-defined surplus of either negative (electrons) or positive (holes) carriers, leading to n-type conductivity or p-type conductivity, respectively. n-type conductivity is obtained by doping with donors such as P, As, and Sb, and p-type conductivity is obtained with acceptors such as B, Al, Ga, and In. Donor or acceptor atoms must be on substitutional sites in the single-crystal silicon lattice to be electrically active, i.e. to form levels close to the conduction band or close to the valence band, respectively. The particular cases of N in Si and C are discussed in the context of ion implantation.

The first, now conventional techniques used to dope a wafer or mask-defined areas of a wafer were based on the diffusion of dopants from the specimen surface, used as the source, into the bulk in long-lasting high-temperature furnace processes. The dopants are already electrically active at the end of the diffusion process. The now conventional implantation of high-energy dopant ions, introduced at the end of the 1960s is performed using high-voltage accelerators. It is a comparatively short process. However, the dopant ions do not necessarily come to rest at substitutional sites. Thus a short high-temperature process must follow to electrically activate the dopants and also to anneal the radiation damage caused by the ion implantation, so we do not really escape processes which involve high-temperature induced changes of sites, i.e. diffusion. Annealing can be done in a furnace, or by laser or electron beam irradiation, for example. Other methods of doping include doping during oxidation, CVD, epitaxy, or MBE, and doping by neutron transmutation. Thus the primary energies of the dopants range from approximately 0.1 eV to the high-MeV level, a range of 10 orders of magnitude. The continuing interest in silicon is demonstrated by the proceedings of large-scale conferences on the occasion of the 50th anniversary of silicon (see e.g. [1]).

We had to learn that extraordinary precautions have to be applied concerning cleanliness in high-temperature processes to avoid deleterious con-

tamination of the specimens due to the carrier lifetime killers gold and iron. Why were the Si wafers suddenly contaminated with traces of Au? It was finally found out that the supplier of the solvent used to clean the wafers had improved the production process of the solvent by introducing a gold catalyst. We and our distant partners in diffusion had to learn that annealing low-dose B-implanted Si specimens in standard furnaces may give odd results owing to cross-contamination.

11.2 High-Temperature Doping by Diffusion

Sources for doping Si specimens from the gase phase include B_2H_6 , PH_3 , and AsH_3 . The process temperatures range from 800°C to 1200°C [13]. Diffusion of a particle is observed when there is a gradient of the chemical potential μ_A . The chemical potential μ_A is the real driving force, besides other forces such as local electrical fields due to states in the crystal lattice; see the detailed discussions in [2] and [3]. Assume that the gradient of the chemical potential in the x direction is caused by a concentration gradient $\partial C/\partial x$ of a species and that the concentration of the diffusing species is low enough, say $C < 10^{18} \text{ cm}^{-3}$, that the interaction between these species can be neglected; then the rate F_x of transfer of the diffusing substance per unit area of a section in the x direction inside an ideal solid may be described by Fick's first law [4],

$$F_x = -D \frac{\partial C}{\partial x}, \quad (11.1)$$

where $C(x)$ is the concentration of the species, and D is the diffusion coefficient. D of random, activated processes under thermodynamic equilibrium conditions is often governed by the Arrhenius equation

$$D = D_0 \exp\{-H/(kT)\}, \quad (11.2)$$

where D_0 is the preexponential factor, H is the activation enthalpy, and k is Boltzmann's constant. The relevant continuity equation, Fick's second law, turns out to be

$$\frac{\partial C(x, t)}{\partial t} = D \frac{\partial^2 C(x, t)}{\partial x^2}. \quad (11.3)$$

In practice the concentration profile is often approximated by the assumption that it falls to some value at a depth l , e.g. 1% of the maximum at the specimen surface $x = 0$, leading to a characteristic diffusion length l defined by

$$\exp\{-l^2/(4Dt)\} = 0.01 \quad \text{or approximately} \quad l \sim 2(Dt)^{1/2}. \quad (11.4)$$

The quantity $(Dt)^{1/2}$ may be used as a rule of thumb or as something like a natural unit of length for diffusion.

Theoretical treatments of many specific examples are presented in [4]. The appearance of *error functions* and *error function complements* is characteristic of the solutions of almost all such examples, e.g., diffusion during an epitaxial process with an interface moving with a velocity v ,

$$C(x, t) = C_0/2[\operatorname{erfc}\{(x - vt)/(2(Dt)^{1/2})\} + \exp\{vx/D\}\operatorname{erf}\{(x + vt)/(2(Dt)^{1/2})\}]. \quad (11.5)$$

Diffusion processes in an electric field \mathbf{E} , a mechanical field, or a gravitational field give results of the following form. For an electric field \mathbf{E} , the important quantity is $\mu\mathbf{E}$; writing $|\mathbf{E}| = E$, the result is

$$C(x, t) = C_0/2[\operatorname{erfc}\{(x - \mu Et)/(2(Dt)^{1/2})\} + \exp\{\mu Ex/D\}\operatorname{erf}\{(x + \mu Et)/(2(Dt)^{1/2})\}] \quad (11.6)$$

The diffusion of electrons or holes can be treated in an analogous way, leading to the famous *Einstein relation*: e.g. for holes,

$$\mu_h kT = eD_h. \quad (10.6a)$$

11.3 Defects and Diffusion Mechanisms

The “perfect” crystal lattice is not perfect at temperatures $T > 0$ K. Boltzmann is responsible for that. The description of diffusion on the basis of Fick’s law is phenomenological. However, diffusion is really an atomic process. Thus a crystal lattice in which diffusion takes place cannot be an ideally perfect lattice. However, a crystal lattice at an absolute temperature $T > 0$ K contains intrinsic defects, i.e. vacancies in a finite concentration, as follows from Boltzmann:

$$N_v = \exp\{\Delta S/k\} \exp\{-\Delta H/(kT)\}. \quad (11.7)$$

Here H is the formation enthalpy of a vacancy and S is the entropy; for silicon, the values are $\Delta S \sim 4 \text{ eV K}^{-1}$ and $\Delta H = 2 \text{ eV}$. These defects, or point defects, have a radius of action of only a few atomic spacings. Tracer methods, Rutherford backscattering, etc. can be applied to analyze diffusion processes.

Atomic and quantum mechanical points of view of the rates of chemical reactions and diffusion, and the influence of the shape of the energy barriers involved and the activation energies were treated theoretically in [5], long before they became relevant for silicon. However, changes of sites of defects can also be treated dynamically, considering the displacements as due to the superposition of phonons. Diffusion and the lattice defects involved are discussed in detail in [6, 7], for example.

Diffusion by means of single vacancies, for instance, dominates over diffusion by divacancies. Diffusion of substitutional atoms at elevated temperatures involves the intrinsic defects which are already present in thermal equilibrium. Self-interstitials, for instance, are intrinsic atomic defects. Thus the diffusivities of the intrinsic defects are involved. Diffusion involves the energy E_n for breaking a bond and the enthalpy H_f for forming a Schottky defect or a vacancy. Metals such as Au are extrinsic atomic defects, i.e. mainly foreign interstitial atoms/interstices or substitutional atoms. Au diffuses by an *interstitial mechanism*, finding an excessive surplus of empty interstitial sites. Thus this diffusion process is governed by the migration enthalpy H_m only; the formation enthalpy H_f does not enter the energy balance. Consequently, the diffusion coefficient D_i of interstitials is 5 to 6 orders of magnitude larger than D_{sub} for substitutional diffusion. High-temperature (i.e. near the melting temperature T_m) Si self-diffusion and group III diffusion have been proposed to occur by an interstitialcy mechanism [8], involving “liquidized” self-interstitials [9]. Two types of swirl defects, type A and type B have been observed and were controversially discussed to be based on vacancies or on interstitials. Electron microscopical analysis of such silicon samples showed that the type A swirl defect consisted of dislocation loops formed by aggregation of interstitials during cooling down from T_m [10], and that the small type B defect [11] was composed of vacancy globules inducing almost no lattice strain. The more or less undisturbed coexistence of interstitials and vacancies, which do not annihilate each other, is interpreted as being due to a large energy barrier (or entropy barrier) to vacancy–interstitial recombination see e.g. [12].

The interstitialcy mechanism is defined as an interstitial atom exchanging sites with atoms on lattice sites, pushing them into neighboring interstices (foreign interstitials). At lower temperatures it is claimed that Si self-diffusion changes to a vacancy mechanism. Group V diffusion occurs via a vacancy mechanism. The large preexponential factor D_0 of the vacancy self-diffusion mechanism is due to a strong inward relaxation of vacancies, with a small, charge-state-dependent activation volume V_a for self-diffusion. Note that most of the point defects in semiconductors may have more than one electric charge state. The process, by which the sites change, depends on the charge state. External stresses may cause relaxation effects. The *crowdion* mechanism, one more atom in the lattice than the lattice permits, needs only a weak movement to bring the crowdion from one site to another one. The *ring* mechanism, however, is less probable.

11.3.1 Lattice Defects, Diffusion, and Gettering

Extended defects such as dislocations play a major role in diffusion. Screw and edge dislocations, dislocation loops, dislocation networks, etc., are the consequence of a high density of atomic defects and of some inner stress,

caused for example by interstitials. They can also be a consequence of vacancies, which agglomerate to form divacancies, diinterstitials, trivacancies, etc., up to dislocation loops, or of precipitates of dopants, e.g. P, i.e. of impurity atoms with an effective atomic radius different from the host lattice. The Si single-crystal lattice is expanded by Al, Bi, Ga, Ge, and In and contracted by B, C, N, and P. The quantity pertinent to the effect is the misfit parameter $\epsilon = (\text{atomic radius of Si} - \text{atomic radius of impurity})/(\text{atomic radius of Si})$. Note in this context the dependence of the diffusion coefficient on the type of doping of the host material.

Dislocations are formed until the inner stress disappears. The diffusion rate along dislocations is considerably higher, owing to lower activation energies, than in the normal lattice. This property is utilized to getter obtrusive impurities such as Au or Fe at dislocations, i.e. to *clean* the solid. Enhanced diffusion is observed at grain boundaries under oxidizing layers. Segregation of dopants occurs at the SiO_2/Si interface, e.g. piledown in the case of B and pileup in the case of P and As. Dislocations can be made visible as etch pits on the specimen surface using a Sirtl etch, by electron microscopy, by infrared microscopy, and by other methods.

11.4 Ion Implantation

In the infancy of electronics, semiconductor devices were fabricated on single-crystal silicon wafers by diffusing [4] dopants into areas defined by masks. Some characteristics of diffusion processes and techniques are (i) a concentration maximum of dopants at the specimen surface; (ii) extended lateral diffusion under the mask edges, not tolerable in modern submicrometer device structures; (iii) difficulties in precisely obtaining predetermined doping levels, mainly at shallow penetration depths, and in obtaining a planar, smooth diffusion front, owing to the influence of lattice defects of all kinds; and (iv) the huge efforts required to obtain clean conditions.

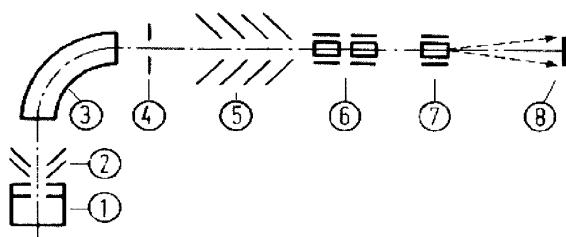


Fig. 11.1. Ion implanter. (1) Ion source, (2) immersion lens, (3) mass separator magnet, (4) exit slit of mass separator, (5) electrostatic accelerator, (6) electrostatic quadrupole focusing lenses, (7) electric scanner, (8) target

At the end of the 1960s the industry introduced ion implantation (Fig. 11.1), which was considered to just fit the new planar techniques, with the hope of overcoming the drawbacks of diffusion without being confronted with too many new ones [13, 51–53]. We learned very soon that ion implantation involved extremely complex problems in all respects compared with diffusion. People, deep-rooted in diffusion processes, were sceptical that devices processed with ion implantation would ever work. The general, inherent radiation damage was said to be prohibitive. It was clear from previous work on radiation damage in metals that ion implantation had to be followed by a second and no less complex process of the same quality, viz. a high-temperature annealing step to induce atomic changes of sites into more stable ones, to restructure the crystal lattice, and to move the dopant atoms into substitutional sites. What is that other than diffusion? We have short-distance diffusion here and long-distance diffusion there. It looked as if we had cured one evil with an even worse evil. Nevertheless, ion implantation was a big step ahead. Knowledge in almost all fields of physics turned out to be very helpful. In spite of all the initial turbulence, ion implantation very fast became a standard technique to fabricate microelectronic devices.

This enthusiasm caused the maximum of the concentration profile of the dopants to be situated not at the target surface but inside the target, and the more distant from the target surface the higher the energy of the dopant ions (in the keV range and up). The different types of concentration profiles are shown in Fig. 11.2. Precise control of the impurity profile opened up the possibility of implanting predetermined complex, so-called tailored profiles by varying only the ion energy and ion dose, to fabricate varactors [53], for example. The leading early application of ion implantation, however, was to implant into the channels of MOS planar transistors to obtain a precise threshold voltage shift, essential for proper device performance. The possibility of implanting extremely flat concentration profiles simply by reducing the ion energy, using molecular ions, or implanting at higher energies but through a thin auxiliary surface layer of SiO_2 opened up new fields of device design.

A small fraction of the ions are reflected at the target surface, accompanied by surface sputtering. Most of the ions penetrate the wafer surface, are slowed down along a polygonal path because of scattering by the atoms of the host lattice, lose their energy, and come statistically to rest in the interior of the bulk (Fig. 11.2). The lateral spread of the penetrating ion beam, defined by a mask of the proper thickness, is small compared with the lateral spread of dopants observed in standard diffusion processes. Heavy ions show a smaller lateral spread than do light ones.

Atoms, both lattice atoms and implanted dopants, are displaced from their original sites owing to impacts with the high-energy ions. A defective lattice containing, for example, vacancies, interstitials, and agglomerates of them results. The host material can even transmute into an amorphous-like state if heavy ions are implanted with a dose above a certain critical value.

When extremely high-energy ions are implanted the amorphous layer may be sandwiched between a very thin, quite imperfect single-crystal surface layer and the unaffected single-crystal bulk of the substrate (see Fig. 11.5). The maximum of the radiation damage concentration profile is usually closer to the target surface than the maximum of the dopant profile. We were surprised that the general rule that the maximum concentration of defects is found closer to the surface than the maximum concentration of dopants may be broken: lattice vacancies are observed much deeper in a Si specimen than the doping profile of implanted As ions.

The dopant concentration is set via an electrical measurement of the integral ion current or fluence, carefully avoiding falsification due to secondary emission of charged particles from the target surface and from the electrodes and apertures of the ion accelerator, and falsification due to influx of neutralized high-energy dopants. The final implantation result may be influenced erratically by localized atomic-range (primary-zone) high “temperatures” and macroscopic temperature rises which take place in the thin implanted layer during the implantation process; this occurs mainly at high ion fluences. Simultaneous annealing may be observed. In order to keep the damage level permanently low, this effect can be increased by externally heating the wafer.

An auxiliary top layer of SiO_2 serves several purposes. This layer protects the wafer from contamination with surface impurities due to knock-on implantation during the implantation process. The maximum energy transfer in a relativistic, elastic, central, head-on collision is $\Delta E_{max,rel} = 4E\{m_0 M(1 - \beta^2)\}^{1/2}/\{m_0 + M(1 - \beta^2)^{1/2}\}^2$, where E is the initial energy of the moving ion, with rest mass m_0 . Naturally, impurities on the SiO_2 surface are also hit by the dopant ions to be implanted and are pushed into the interior of the wafer, but they come to rest in the auxiliary SiO_2 layer. Later on, this SiO_2 layer is etched away together with the impurities in it. Note that

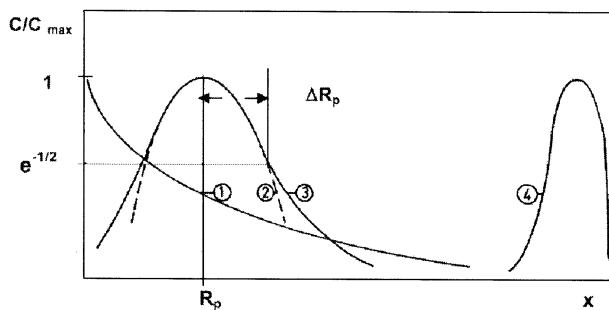


Fig. 11.2. Sketch of normalized concentration profiles $C(x)/C_{max}$ of impurities in a solid: (1) diffusion, (2) Gaussian profile, (3) random implantation, (4) channeling; R_p is the mean projected range, and ΔR_p is the standard deviation

there is a mixing effect between SiO_2 and Si at the SiO_2/Si interface, which may become quite diffuse.

Surface contamination originates, for example, from residues from cleaning procedures or from deposits due to sputtering of the electrodes or apertures of the implanter. After the start of the implantation, atoms, molecules, and radicals of various types sputtered away from the wafer surface are observed in the residual gas during the first few milliseconds. The type and quantity of the contamination can be measured in the target chamber with a high-resolution, fast residual-gas mass analyzer. Part of this contamination of the “clean” wafers can be reduced substantially by keeping the wafers in a high-vacuum airlock for some time before transferring them into the target chamber, which is operated with oil-free cryogenic pumps.

The knock-on effect on predeposited thin surface layers of dopants can be utilized to form extremely thin doped surface layers or to form very good ohmic contacts.

11.5 The Special Case of Silicon, Nitrogen, Carbon, and Dielectrics

The principal dopants of silicon are the very first, lightest elements of the Group III and Group V columns of the periodic table of elements. The exception to the rule concerns nitrogen, which shows only a minimal diffusion rate even at elevated temperatures, and an insignificant activation ratio of only a few percent, if at all [14–20]. The data found in the literature are distinguished by quite a large spread. The diffusion coefficient of atomic N is claimed to be $D = 0.87 \exp\{-3.29/kT\}$, the activation energy $E_{a,N} = 3.29 \text{ eV}$. The activation energy E_{a,N_2} of molecular N_2 is 2.6 eV . Nitrogen behaves as neutral in the Si:N_2 state, the diffusion state. Thus doping by diffusion of nitrogen can be neglected, but it is observed with ion implantation [17], where nitrogen is in the atomic state from the outset. This means that ion implantation of nitrogen is tailored just to fit the analysis of the special case of nitrogen. The electrical activation is measured to be up to 5%, i.e. 5% of the implanted nitrogen atoms are in a substitutional position and 95% in an interstitial position or in other states, within the limits of an N-dose range between $1 \times 10^{13} \text{ cm}^{-2}$ and $5 \times 10^{16} \text{ cm}^{-2}$ [16, 17, 20]. A maximum electron concentration of $2 \times 10^{19} \text{ cm}^{-3}$ has been found [16, 17]. The solubility limit of N in solid silicon is stated to be $4.5 \times 10^{15} \text{ cm}^{-3}$ [16]. The donor levels are claimed to be at 0.02 eV when the Si is in the degenerate state, but at 0.042 up to 0.58 eV under normal conditions. Above a critical concentration of $2 \times 10^{19} \text{ N/cm}^3$, single nitrogen atoms are observed to form bonds with the silicon and can nucleate to form the dielectric Si_3N_4 phase [20]. These effects are attributed to the covalent radius of nitrogen of 0.07 nm , which is much smaller than the covalent radius of silicon of 0.17 nm [14, 20]. It has been claimed that the attribution of the effect to the large s–p energy

Table 11.1. Properties of SiO_2 and Si_3N_4 see. e.g. [25]

Property*	SiO_2	Si_3N_4
Crystal structure	Amorphous for most IC applications	Hexagonal,** amorphous for most IC applications
Lattice constant (nm)		0.75**
Density (g cm^{-3})	2.2–2.65	3.44**
Melting point ($^{\circ}\text{C}$)	1700	1900
Thermal conductivity (W/cm K)	0.014	0.185 (0.018)
Energy gap (eV) at 300 K	8	4.7
Index of refraction at 0.5 μm	1.46	2.0
Relative permittivity	3.9	7.5
Electrical resistivity (Ωcm)	$> 10^{16}$	$10^7\text{--}10^{17}$

* The properties depend strongly on the preparation processes and on traces of impurities. Silicon nitride usually corresponds more to the irregular Si_xN_y form than to the stoichiometric form.

splitting, leading to an absence of sp^3 -bonded N_4 sites, is wrong but that the small size of nitrogen, expressed using the concept of nonbonded radii, makes it impossible to accommodate four silicon atoms at the normal bond length of 0.175 nm [19]. Further, the deep donor level at 0.58 eV is attributed to isolated N donors $\langle 111 \rangle$ trigonally distorted owing to the Jahn–Teller effect [14, 18]. A similar inconsistency concerning carbon is observed when one tries to form the analogous compound C_3N_4 [21] by ion implantation of nitrogen into carbon or carbon/silicon targets. The maximum atomic concentration of nitrogen reached only a value of up to 45% under certain complex conditions, instead of the necessary 57% [22]. Tight-binding model calculations show an extremely small electron conductivity in this graphitic C_xN_y [15]. Large-scale formation of dielectric C_3N_4 was not observed.

Dielectric layers of Si_3N_4 , [23, 24], SiO_2 , and $\text{Si}_x\text{N}_y\text{O}_z$, are formed deep inside silicon wafers by high-dose, high-energy implantation of nitrogen and/or oxygen. Silicon oxynitride layers are attractive because it is expected that the different suitable properties of SiO_2 and Si_3N_4 [25] can be merged so as to be attractive in microelectronics [26], optoelectronics, and solar cells, [27], see also Table 11.1. Some interesting properties are the amorphous structure, barrier properties, mechanical hardness, dielectric function, loss factor, IR absorption, radiation hardness, and electrical breakdown strength. The

mechanism of the electrical breakdown observed in ultrathin SiO_2 layers, however, seems not to be clear yet. Efforts to deposit thin, structured surface layers of Si_3N_4 or $\text{Si}_x\text{N}_y\text{O}_z$ on Si by laser-chemical photomethods have not achieved success without problems; for example, reproducibility, necessary in industrial applications, is a problem. Thus the ion implantation of nitrogen becomes a real alternative to thermal diffusion.

11.6 Implantation Profiles

Three basic types of concentration profiles have to be considered (Fig. 11.2): (i) An almost Gaussian concentration distribution is observed when the wafer is irradiated at 7° from a crystallographic orientation, a so-called random direction for the ion beam. The distance x of the concentration maximum measured from the wafer surface is called the mean projected range $R_p(x)$. (ii) When the wafer is oriented such that the ions (beam aperture $< 1^\circ$) impinge on the wafer in an exact crystallographic orientation, e.g. parallel to the very open [110] axis of silicon, a relatively sharp concentration peak of dopants appears owing to *axial channeling*, much deeper in the wafer than is observed in case (i). The channeling effect is based on the reduced energy losses caused by reduced scattering. A weak case (ii) due to rechanneling effects can be also observed in practice. Profiles of type (ii) are difficult to reproduce. They are avoided under fabrication conditions, and profiles of type (i) are preferred. Planar channeling, also observed, is a very small channeling effect in lattice planes. Channeling can be prevented by preimplanting sufficiently highly energetic silicon ions to amorphize the relevant silicon layer or by additionally depositing before the implantation a thin amorphous SiO_2 layer, for example, which serves to spread the ion beam by scattering to such an extent that the channeling effect is minimized. However, a rechanneling effect cannot be avoided even when an auxiliary amorphous SiO_2 layer is utilized. The oxide layer is usually etched away after the implantation process is finished.

The main relevant energy loss processes are inelastic electronic interaction, i.e. excitation of bond shell electrons and free target electrons, and elastic Coulomb interactions between screened nuclear charges, i.e. the nuclear stopping. The stopping power $-dE/dx$ is defined by these processes. The main total specific energy loss becomes

$$dE/dx = (dE/dx)_e + (dE/dx)_n, \quad (11.8)$$

i.e. the sum of electronic-type and nuclear-type energy losses. Other types of losses, such as inelastic stopping by nuclear interaction and losses at the very lowest energy range, say below 10 eV, when the dopant is almost coming to rest can usually be neglected. In that energy range chemical binding forces should be taken into account and quantum mechanical calculations done. Energy loss due to charge exchange may also be neglected. Figure 11.3 shows the

dependence of the electronic and nuclear stopping powers on the ion energy. The nuclear stopping power starts to dominate at quite low ion energies and is mainly responsible for lateral scattering and hence for the lateral spread.

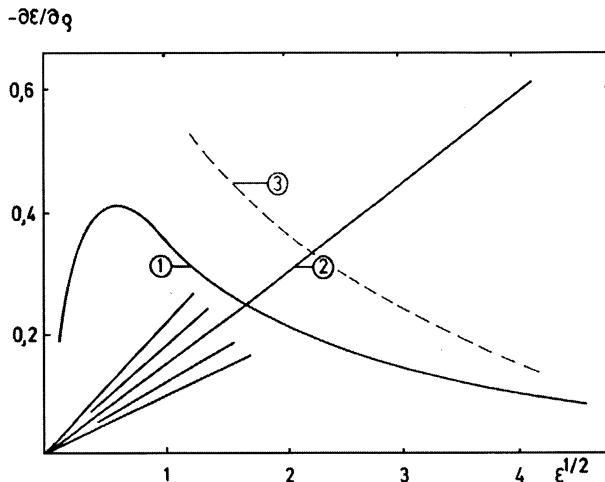


Fig. 11.3. Schematic representation of the reduced specific energy loss versus reduced energy: (1) nuclear (Thomas–Fermi potential), (2) electronic, (3) nuclear (unscreened Coulomb potential, Rutherford scattering)

The calculation of the electronic stopping power depends on the atomic model used. The LSS theory (Lindhard, Scharff, and Schiøtt), for instance, is an approximation assuming the atomic Thomas–Fermi model (self-consistent field method) to be valid (e.g. [28]). However, profile asymmetry, tails of complex origin in the direction towards the bulk, etc. have to be considered. Approximative values of R_p and the standard deviation ΔR_p have been calculated as functions of the initial ion energy E and compiled in tables (e.g. [29]) that are useful and sufficient for practical purposes, say for industrial applications. Most of the theoretical values corresponded well enough with the experimental values that could be determined at the time. Methods of analysis included measurement of the change of the electrical conductivity of electrically activated dopants when the implanted layer was successively etched off in steps of, say, 20 nm thickness; C – V measurements of MOS capacitors; and direct particle analysis by electron spectroscopy, SIMS, secondary ion mass spectrometry, and RBS, Rutherford back scattering analysis. It was an interesting experience to see presented high-resolution SIMS profile measurements demonstrating almost perfect Gaussian implantation profiles and, two years later, SIMS profiles measured on identical specimens with the same high resolution but demonstrating pronounced tails.

In many cases the concept of a mean projected range R_p and a Gaussian concentration profile $N(x)$ may still be useful today:

$$N(x) = N_{max} \exp\{-(x - R_p)^2/[2(\Delta R_p)^2]\}. \quad (11.9)$$

The width of the Gaussian profile at the normalized value $N/N_{max} = 1/e^{1/2}$ determines the value of the standard deviation ΔR_p . By integrating (11.9), the value of N_{max} can be calculated approximately when the total dose N is known, and vice versa:

$$N_{max} = N/\{(2\pi)^{1/2} \Delta R_p\} \sim 0.4 N/\Delta R_p. \quad (11.10)$$

Computer simulation programs, e.g. [30], simplify the work. However, a proper simulation program must be selected when spatial and time-dependent variations of the density of the implanted layer have to be taken into account. Such conditions may exist under high-dose implantation of O or N ions in Si to produce insulating dielectric layers or diffusion barriers of SiO_2 , $\text{Si}_x\text{N}_y\text{O}_z$ or Si_3N_4 [24–26]. When complex concentration profiles, say tailored ones, are required, the exact shape of the profile, including deviations at the flanks, has to be determined. The simulations can be verified experimentally using NRA, nuclear reaction analysis or RBS, for instance.

It was claimed, the first time ion implantation was used, that ion implantation reproduced the mask edges in an ideal way, in contrast to diffusion. This claim was corrected theoretically and experimentally [31], when a substantial lateral spread of the ion beam, due to scattering and outdiffusion of the dopants due to the high-temperature process used to activate the dopants and to anneal the radiation damage, was demonstrated.

11.7 Sputtering and Profiles

The sputtering of target surface atoms due to impinging high-energy ions affects the concentration profile of the implanted ions at high sputtering yields, and may even lead to a saturation profile with the maximum at the specimen surface when the number of impinging ions reaches the same level as the number of sputtered dopants. A strong effect on implantation of Li ions in GaAs was reported very early [32]. This effect should be kept in mind when silicon changes its structure during high-dose implantation, for example of N to produce Si_3N_4 [24]. Mysterious-looking effects during formation of compounds by ion implantation may be attributed with caution to a suddenly changing sputtering yield when phase changes, caused for example by chemical reactions, are involved. Thus, under certain conditions, even ion implantation exhibits limited concentration maxima at the target surface like those obtained by diffusion, but these effects are induced by sputtering [33]. When the sputtering yield is very small, the implanted dose of dopants can be higher than their solubility limit observed in diffusion.

11.8 Radiation-Induced Defects and States, Surface States, and Interface States

The high-energy dopant ions slow down in the bulk of the target along a polygon-shaped path, losing energy and momentum to the crystal lattice owing to collisions with the atoms of the host lattice. When the lost energy is high enough the collision partner of the host lattice can be displaced, leaving behind a vacancy, resulting in a vacancy close to a self-interstitial, known as a Frenkel pair. The knock-on particle, for its part, may cause radiation damage as well when the energy transferred to it is high enough to do so, resulting in secondary radiation defects and finally a whole collision cascade [6, 7]. A big part of the energy which the dopant ion loses is dissipated at the end of its path, causing a region of heavily damaged material. There the “temperature” locally increases to extremely high values far away from thermodynamic equilibrium. This “temperature” decreases substantially over short distances of a few lattice periods. The lattice becomes thermalized. The defects may change sites, depending on the activation energy involved, during this period. Agglomerates of atomic defects and complicated complexes of defects may be formed during implantation when the temperature of the specimen becomes too high. Excess point defects such as self-interstitials may agglomerate into extended defects, e.g. stacking faults and dislocation loops (see e.g. [34]). Amorphous phases formed by implanting extremely high doses become important for reordering of the implanted layer through solid-state epitaxial regrowth during annealing in order to obtain a minimum of residual defects.

Such defects may be sinks for dopant atoms and hence eliminate them for the purposes of doping. Such defects may cause deep levels inside the bandgap. The deep levels can act as generation/recombination centers, reducing the lifetime of the electrical carriers. The defects increase the electrical resistivity by scattering the carriers.

A strong point of Si technology is the natural formation of SiO_2 , a fortunate, unique gift of nature. Consequently, Si specimens can easily be provided with an SiO_2 layer, as one of the principal dielectric layers, by simple thermal oxidation, in the LOCOS, local oxidation of silicon process, etc. SiO_2 layers serve as implantation or diffusion masks, amorphous oxides to reduce channelling, protective layers to avoid knock-on implantation of surface impurities, passivation, insulators with high electric breakdown fields, etc. A particular critical point is the SiO_2/Si interface and its complex structure. The interface states involved include P_b -centers, with the structure $\bullet\text{Si}\equiv\text{Si}_3$ or $\bullet\text{Si}\equiv\text{Si}_2\text{O}$, containing a dangling bond [35]. The structure of the P_b -center is amphoteric, with two states and thus two energy levels, a donor and an acceptor level. The interface state density distribution D_{it} shows two peaks [36]. This center is related to the origin of leakage currents observed in devices, mainly along implantation-mask edges [37]. The dangling bonds are neutralized using hydrogen, but it has also been reported that a more stable configuration is obtained using deuterium [38]. Note in this connection that the SiO_2/Si in-

terface may become smeared out, undergoing a mixing effect due to knock-on implantation.

However, deficiencies of SiO_2 , such as limited chemical and physical stability and an inconvenient dielectric function, became evident in advanced device fabrication and initiated the introduction of other dielectrics, mainly Si_3N_4 , Si_xN_y , and $\text{Si}_x\text{N}_y\text{O}_z$ formed by deposition techniques or by ion implantation [24]. Of course, the nitride compounds of silicon are also burdened with ion-implantation-induced defects and natural defects. The $\text{Si}_3\text{N}_4/\text{Si}$ interface quality of Si_3N_4 is usually inferior to that of SiO_2 . The problem can be solved by inserting an $\text{Si}_x\text{N}_y\text{O}_z$ transition layer. The defects on the atomic scale include traps, dangling bonds, bond angle distortions, and current-induced low-lying states. Such defects are often quite annealing-resistant. Besides the already noted P_b -center there is the neutral K° , the $\bullet\text{Si} \equiv \text{N}_3$ defect, which has an unpaired electron wave function of 21% 3s character and 49% 3p character. The K^+ and K^- defects are paramagnetic. Some other defects are the traps D^+ , D° , and D^- , and the N_4 and N_2 states with their different charge states, the VAPs, valence alternation pairs. The interface state density D_{it} changes with the annealing temperature. The radiation hardness is reported to increase with decreasing Si_3N_4 layer thickness. Some methods of analysis are ESR, NMR, and MAS NMR, magic-angle spinning nuclear magnetic resonance measurements. An extensive discussion can be found in [39].

11.9 Annealing of Ion-Implanted Specimens

11.9.1 Furnace Annealing

During the first few years of ion implantation, the electrical activation of the implanted dopants and the annealing of the damaged host lattice by inducing changes of sites were exclusively performed by applying long-lasting (10 min and more) conventional furnace processes at temperatures around 450°C in protective inert-gas atmospheres of N_2 or N_2/H_2 mixtures to remain compatible with existing techniques and to allow seamless integration into established industrial production processes. However, only incomplete recovery of the damaged target lattice and of deep levels was obtained, and not a real, complete activation of the dopants. Many papers addressed this problem in the earliest international conferences, claiming that temperatures of at least 650°C and annealing times between 10 and 30 min were necessary [40]. Today the temperature range goes up to 1200°C to eliminate also primary and secondary defects which have high activation energies, up to several electron volts. Recrystallization at the proper temperature is said to occur epitaxially, passing through, for example, a state of defect clusters or dislocation loops, which finally may be dissolved. More defects, however, remain in (111) Si wafers than in (100) Si wafers after epitaxial regrowth. The remaining resid-

ual defects can only be eliminated by remelting the material, a process which cannot be performed when fabricating complex electronic devices.

The annealing of high-dose implanted amorphous Si substrates is not necessarily a straightforward process. Implantation of boron or BF_2 molecules may serve as an example. Three annealing stages are observed, one up to approximately 500°C , a second one between 500°C and 650°C showing pronounced reverse annealing, and a third stage above 650°C [41]. Point defects are said to agglomerate between 600°C and 800°C , forming dislocations. Thus an optimum annealing temperature is found for certain conditions at approximately 850°C [48], owing to the influence of the neighboring surface acting as an inexhaustible sink for defects but also acting as a source. A minimum of left-over defects is observed when amorphous silicon is annealed by solid-state epitaxial regrowth. However, hairpin dislocations are observed to be formed more often in (111) wafers than in (100) wafers. Good annealing can easily be achieved when doses either much below the amorphization dose of the specimen material or much above are implanted. Annealing becomes problematic when doses in between are implanted. High-dose implantation may cause two parallel damage zones exhibiting imperfect annealing (see Fig. 11.5e). p-n junctions should preferably not be put just inside such a defective zone.

A proper slow ramping would ensure that the specimen reached almost thermodynamic equilibrium conditions after cooling down to room temperatures. However, a finite concentration of defects, say of vacancies, is always found even in a perfectly annealed crystal at a temperature $T > 0\text{ K}$ according to the laws of thermodynamics, i.e. according to the statistics of the equilibrium state, because the entropy increases when a vacancy is formed. The deficiency of the long-lasting furnace processes concerns particularly the lateral outdiffusion of the implanted dopants at mask-defined zones or, generally, a change of the original concentration profile [4]. A pronounced outdiffusion from the zones of the maximum concentration may take place at implantation doses with a peak concentration above the solubility limit and lead to a square-well-shaped concentration profile, particularly when As is implanted into Si. Under certain conditions, even diffusion in the direction of higher implantation concentration can be observed under the influence of the gradient of the chemical potential or when chemical reactions take place. The usual small diffusion coefficient at low implantation doses can become very high when the host material has become amorphous, along dislocations or grain boundaries formed during the annealing. Such effects are prohibitive for devices particularly with structures in the submicrometer range. Further, segregation is observed at boundaries between two materials, e.g. single-crystal silicon and amorphous SiO_2 . The segregation coefficient is the ratio between the concentrations of an impurity at the one side of the boundary and at the other side when thermodynamic equilibrium is reached. This situation was the motive for looking for new annealing techniques which permit shorter annealing periods.

11.9.2 Electron Beam, Laser Beam, and Rapid Thermal Annealing

Some relevant techniques were already available, and new ones could be added to overcome the drawbacks of furnace annealing. However, the devil is in the detail. This concerns less the scientific field, where such new techniques may even be the sheet anchor, but may restrict their application in industrial production to particular cases. Each of these techniques has qualities which strike the eye, but also prohibitive ones. The various techniques will not be discussed separately but will be contrasted with each other on the critical points.

Pulsed laser processing (only one single pulse, irradiating the whole wafer at once) is one of the most extreme techniques, lasting for nanosecond periods down to process periods in the femtosecond range. It is fascinating that annealing processes can work over a range of durations which covers more than 10^{15} orders of magnitude. Continuous-wave (cw) laser annealing, electron beam annealing, intense incoherent lamp irradiation, etc., belong to the millisecond up to the multisecond range, the so-called rapid thermal annealing range. All these methods have the advantage that the dopants become electrically active and large-scale redistribution of dopants is insignificant as long as *solid state conditions* are maintained, i.e. no melting occurs, not even of a thin surface layer. However, the thermodynamic state at the highest temperature reached is frozen in to a certain extent owing to the rapid cooling. The specimen remains in thermodynamic nonequilibrium conditions with, for example, a substantial concentration of atomic defects. Such defects can be neutralized by applying afterwards a moderate-temperature process in a hydrogen-containing atmosphere, as is done during alloying of contacts. For a comprehensive monograph, see [49].

The strong point of short-wavelength pulsed laser annealing – particularly in the ultraviolet range to obtain the smallest penetration depth of the light – is the annealing or alloying of ultrathin (down to the nanometer range) surface layers, which can be defined by imaging or masks on extremely small areas of a wafer. Such a process is interesting, for instance, in research and in the case of customer-designed small industrial production volumes when existing devices are to be modified, for example, by an additional contact, and the finished wafer or chip cannot be heated up as a whole anymore. On the other hand, it can be demonstrated drastically by catastrophic experiments that the coherence of the laser light and its material-dependent absorption may cause serious annealing problems [42]. Such catastrophic experiments can be useful for obtaining basic hints in the shortest possible time and with a minimum of effort. To this end, the laser intensity, for instance, has to be selected such that the temperature becomes just high enough in the implanted, free silicon areas of MOS structures defined by an SiO_2 mask. The reaction at the SiO_2 -covered areas, however, can then be disastrous (see Fig. 11.4) when the thickness d_2 of the SiO_2 layer just meets the condition for an antireflection coating, $d_2 = (2m + 2)\lambda/(4n_2)$, $m = 0, 1, 2, 3, \dots$, where

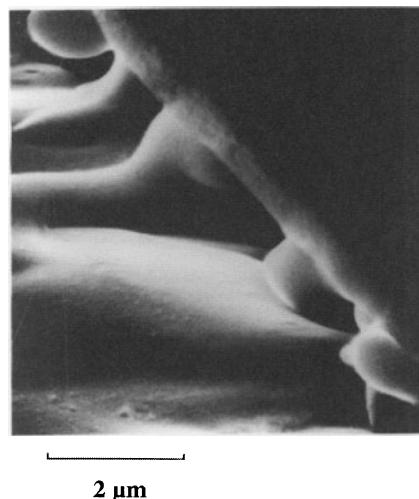


Fig. 11.4. MOS structure pulse-irradiated by a ruby laser (*left*) to anneal an implanted Si area masked by an SiO_2 layer. Part of the dielectric (*left*) has been blown away by evaporating Si owing to the SiO_2 layer thickness just meeting the conditions for an antireflection coating

n_2 is the refractive index, and λ is the laser wavelength. In this case much more energy of the laser light is coupled to the bulk of the silicon under the SiO_2 layer than to the uncovered silicon, where part of the light is reflected. In the irradiated area, the covering SiO_2 mask layer has been blown away by the silicon evaporating under it; the uncovered silicon remains intact. Other critical points should be noted. The absorption of light depends on the wavelength and is generally, in the range applied, much higher for silicon than for SiO_2 , leading to temperatures higher in Si than in SiO_2 and hence to thermo-mechanical stresses at Si/ SiO_2 boundaries. Such stresses may cause leakage currents, mainly at mask edges [37]. The defects responsible can be eliminated in part by a following low-temperature anneal in a hydrogen atmosphere. This means that the various parameters of the laser process have to be made compatible by a careful selection.

Note that equivalent detrimental problems may arise when, for example, photochemical reactions induced by pulsed laser light irradiation of precursor gases are used to deposit Si. The coherence and spatial inhomogeneity of the laser light and, consequently, of the matter in the gas space can cause extremely high local electromagnetic fields such that a local breakdown occurs. Ionization of atoms due to inner-shell electron emission can lead to unexpected chemical reactions with a chaotic distribution. It seems that these effects can dominate multiphoton transitions so as to launch chemical processes. Further, the electric charges can act as nucleation centers in the space of the gas phase. The existence of free electrons can be easily proved and

the amount of these electrons measured using e.g. a primitive ballistic galvanometer [50].

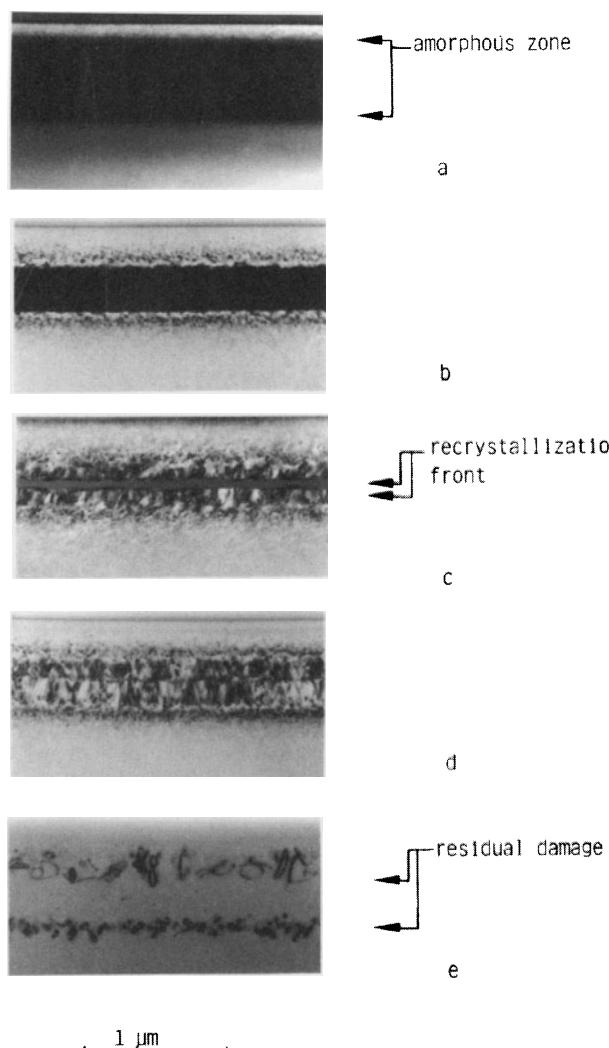


Fig. 11.5. Recrystallization kinetics of a zone amorphized by high-dose, high-energy ion implantation. TEM micrographs of cross-sectional, 760 keV As-ion-implanted Si specimens. (a) as-implanted; the amorphous zone (black) is covered with a thin single-crystal Si layer (white); (b), (c), (d) different annealing steps of scanned-electron-beam-annealed specimens; (e) 30 min furnace annealing

Scanned electron beam annealing can be an alternative to whole-area pulsed laser annealing. The process temperature of the specimen may be controlled using a filament pyrometer. Coherence and interference effects do not exist under conventional process conditions as long as the width of the electron source is large compared to the a in the angular-coherence condition $a 2 \sin \vartheta \ll \lambda$, where ϑ the angle under which the electron source is seen, λ is the de Broglie wavelength of a relativistic electron, $h[2m_0eU(1 + eU/\{2m_0c^2\})]^{-1/2}$; for practical use, $\lambda \sim (1.5/U[\text{volt}])^{1/2}$ [nm].

The energy loss of the electrons in the masking layers does not differ too much from the energy loss in the silicon. Thus the temperatures in the different layers will not be too different, reducing thermo-mechanical stresses. The process times range from milliseconds upwards and hence are much longer than in pulsed laser annealing but are still short enough to minimize diffusion effects. Charging effects in dielectrics may cause problems sometimes, but can be solved. Electron-beam-annealed specimens are in better thermodynamic equilibrium conditions than pulsed-laser-annealed ones owing to the slower ramping during the cooling-down period. Defect-induced leakage currents are observed to be even smaller than those obtained after laser and furnace annealing after the influence of defects, i.e. dangling bonds, is reduced substantially by a subsequent hydrogen treatment. The results of annealing can be improved by applying a scanned line focus, using a magnetic quadrupole lens, for example, instead of a scanned point focus, and orienting the line focus parallel to the flat, i.e. to the crystallographic orientation of the wafer [43].

The recrystallization kinetics and the character of amorphous zones formed by high-dose ion implantation of As ions in silicon can be made visible using an appropriate electron beam annealing. This is shown in Fig. 11.5 by a series of transmission electron microscope (TEM) micrographs taken on cross-sectional specimens [44]. Owing to the surface acting as an inexhaustible sink (or source) for migrating defects, the interface facing the surface recrystallizes much faster than the interface facing the bulk. The recrystallization fronts do not exhibit planar interfaces. The completely annealed specimen shows two very thin zones with a high defect density, one where the two recrystallizing interfaces met and a second where the transition layer between the amorphous zone and the crystalline bulk was situated (see Fig. 11.5). In comparison, if the whole volume is doped by diffusion to the same high level, it is loaded with a high density of defects of all types; such defects are also used for gettering of unwanted impurities such as Au or Fe. However, one has to keep in mind that a cross-sectional sample prepared for TEM micrographs has suffered substantial preparation processes, which may have changed its original structure by energy transfer, owing to the close surfaces, mechanical stresses, reordering, etc.

Scanned-electron-beam-induced liquid-state reordering of polycrystalline silicon layers grown on amorphous SiO_2 layers results in crystalline silicon layers containing silicon crystallites as long as 2 mm, attributed to some sort

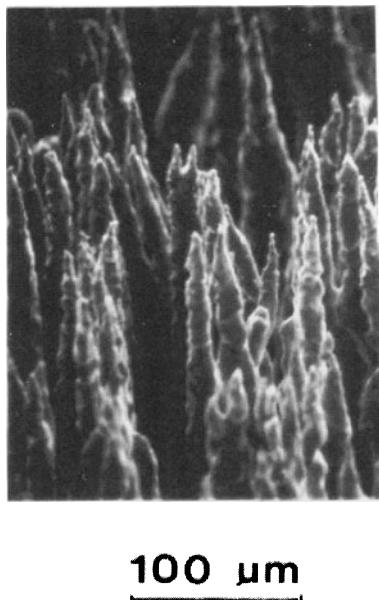


Fig. 11.6. Formation of Si needles from sintered silicon as a result of electron beam recrystallization; REM, reflection electron microscopy micrograph

of pulling effect. The (110) electron diffraction pattern taken in reflection is that of a silicon single crystal [45]. Even the surface of sintered silicon samples containing large flaws can be transformed by scanned electron beam irradiation to a dense crystalline silicon surface layer. Such a surface layer may be interesting for optoelectronics, particularly solar cells. Note, however, that the growth of long needles from a sintered silicon specimen can be observed if the energy transferred from the electron beam to the sintered silicon sample is limited to a level such that the mobility of surface atoms, governed by low activation energies, and the temperature-gradient-dependent diffusion direction become the dominating reordering effects (Fig. 11.6). The relevant energy window is extremely small between epitaxial regrowth and the pulling effect. There is no surprise that the optical appearance of such a shaped surface is a deep black [46]. Correspondingly, growth of silicon nanowiskers has been observed after electron beam annealing of silicon implanted with a high dose of nitrogen, performed in connection with experiments to form Si_3N_4 layers [24, 47].

11.10 Conclusion

We wanted to escape from the considerable inflexibility of obtaining doping profiles by conventional high-temperature diffusion processes, and there-

for introduced ion implantation. However, we did not escape from high-temperature processes similar to those used previously, because they are need to activate the dopant atoms and anneal the radiation damage. Where is the difference, finally? Conventional diffusion is characterized by simultaneous migration of particles and electrical activation. Ion implantation has split this process step into two process steps and has gained a very high flexibility in tailoring profiles, but this has been purchased at the cost of a much more critical high-temperature process than experienced before.

We recognize that the point is not to find the very best process of all, but to find which of the various processes fits best to the fabrication of a particular device. This is shown clearly by trench techniques, a jump into the third dimension to increase the integration density of microelectronic devices, where ion implantation may not be a good solution. However, doping by other advanced techniques such as molecular-beam epitaxy (MBE) will be a solution when devices in the nanometer range are of interest. New techniques concerning quantum dots and structures obtained by self-organization may also be examples of future activities.

References

1. H.R. Huff, H. Tsuya, U. Gösele (eds.): *Semiconductor Silicon 1998*, Proceedings, Vol. 98-1 (Electrochemical Society, Pennington, NJ 1998)
2. J. Bardeen, C. Herring: Diffusion in alloys and the Kirkendall effect. In: W. Shockley, J.H. Hollomon, R.J. Maurer, F. Seitz: *Imperfections in Nearly Perfect Crystals* (Wiley, New York 1952) pp. 261–288; C. Zener: ibid., pp. 289–314
3. J. Crank: *The Mathematics of Diffusion*, 2nd edn (Clarendon Press, Oxford 1985) pp. 212–214
4. J. Crank: *The Mathematics of Diffusion*, 2nd edn (Clarendon Press, Oxford 1985)
5. S. Glasstone, K.J. Laidler, H. Eyring: *The Theory of Rate Processes* (McGraw-Hill, New York 1941)
6. M. Lannoo, J. Bourgoin: *Point Defects in Semiconductors I*, Springer Series in Solid-State Sciences 22 (Springer, Berlin, Heidelberg, New York 1981)
7. J. Bourgoin, M. Lannoo: *Point Defects in Semiconductors II*, Springer Series in Solid-State Sciences 35 (Springer, Berlin, Heidelberg, New York 1983)
8. A. Seeger, K.P. Chik: Diffusion mechanisms and point defects in silicon and germanium., Phys. Stat. Sol. **29**, 455 (1968)
9. M. Werner, H. Mehrer, H.D. Hochheimer: Effects of hydrostatic pressure, temperature, and doping on self-diffusion in germanium. Phys. Rev. B **32**, 3930 (1985)
10. H. Föll, B.O. Kolbesen: Formation and nature of swirl defects in silicon. Appl. Phys. **8**, 319 (1975)
11. S.M. Hu: Defects in silicon substrates. J. Vac. Sci. Technol. **14**, 17 (1977)
12. see e.g. in ref [7] chapter 9, pp. 247–270
13. D. Widmann, H. Mader, H. Friedrich: *Technology of Integrated Circuits*, Springer Series in Advanced Microelectronics Vol. 2 (Springer, Berlin, Heidelberg, New York 2000)

14. H.A. Jahn, E. Teller: Stability of polyatomic molecules in degenerate electronic states. *Proc. Roy. Soc. A* **161**, 220 (1937)
15. H. Bross: private communication (2003)
16. J.B. Mitchell, J. Shewchun, D.A. Thompson: Nitrogen-implanted silicon. II. Electrical properties. *J. Appl. Phys.* **46**, 335 (1975)
17. P.V. Pavlov, E.I. Zorin, D.I. Tetelbaum, A.F. Khokhlov: Nitrogen as dopant in silicon and germanium. *Phys. Stat. Sol. (a)* **35**, 11 (1976)
18. K.L. Brower: Jahn-Teller-distorted nitrogen donor in laser-annealed silicon. *Phys. Rev. Lett.* **44**, 1627 (1980)
19. J. Robertson: Theory of defects in amorphous semiconductors. *J. Non-Cryst. Solids* **77&78**, 37 (1985)
20. N.N. Gerasimenko, V.F. Stas': Buried insulator layer formation by N⁺ implantation. *Nucl. Instrum. Methods Phys. Res. B* **65**, 73 (1992)
21. A.Y. Liu, M.L. Cohen: Structural properties and electronic structure of low-compressibility materials: β -Si₃N₄ and hypothetical β -C₃N₄. *Phys. Rev. B* **41**, 10727 (1990)
22. F. Link: Untersuchungen zur Phasenbildung mittels Ionenimplantation im Si-C-N System. PhD Thesis, J.W. Goethe Universität, Frankfurt am Main (1999)
23. J.A. Borders, W. Beezold: Infrared studies of SiC, Si₃N₄, and SiO₂ formation in ion-implanted silicon. In: *Proc. Second Intern. Conf. on Ion Implantation in Semiconductors*, ed. by I. Ruge, J. Graul, Garmisch-Partenkirchen, Germany (Springer, Berlin, Heidelberg, New York 1971) pp. 241–247
24. A.M. Markwitz: Ionenstrahlsynthese und Analyse von oberflächennahen und vergraben Siliziumnitridschichten. Ph.D. Thesis, J.W. Goethe University, Frankfurt (1994)
25. E.F. Krimmel: Silicon nitride: electronic structure; electrical, magnetic, and optical properties; spectra; analysis. In: *Gmelin Handbook of Inorganic and Organometallic Chemistry, Si, Silicon*, Supplement Vol. B 5b2, 8th edn (Springer, Heidelberg 1997) pp. 1–171
26. E.F. Krimmel: Silicon nitride in microelectronics and solar cells. In: *Gmelin Handbook of Inorganic and Organometallic Chemistry, Si, Silicon*, Supplement Vol. B 5c, 8th edn (Springer, Heidelberg 1991) pp. 1–320
27. R. Hezel: Silicon nitride in microelectronics and solar cells. In: *Gmelin Handbook of Inorganic and Organometallic Chemistry, Si, Silicon*, Supplement Vol. B 5c, 8th edn (Springer, Heidelberg 1991) pp. 321–362
28. J. Lindhard, M. Scharff, H.E. Schiøtt: Range concepts and heavy ion ranges. *Mat. Fys. Medd. Dan. Vid. Selsk.* **33**, No. 14 (1963)
29. W.S. Johnson, J.F. Gibbons: Projected Range Statistics in Semiconductors (Stanford University Book Store, Stanford, CA 1969)
30. J.F. Ziegler, J.P. Biersack, U. Littmark: *The Stopping and Ranges of Ions in Solids* (Academic Press, New York 1983)
31. H. Runge: Distribution of implanted ions under arbitrarily shaped mask edges. *Phys. Stat. Sol. A* **39**, 595 (1977)
32. A.W. Tinsley, W.A. Grant, G. Carter, M.J. Nobes: The retention of bi ions implanted in GaAs. In: *Proc. Second International Conference on Ion Implantation in Semiconductors*, ed. by I. Ruge, J. Graul, Garmisch-Partenkirchen (Springer, Berlin, Heidelberg 1971) pp. 199–204
33. E.F. Krimmel, H. Pfleiderer: Implantation profiles modified by sputtering. *Radiation Effects* **19**, 83 (1973)

34. K.S. Jones, S. Prussin, E.R. Weber: A systematic analysis of defects in ion-implanted silicon. *Appl. Phys. A* **45**, 1 (1988)
35. P.J. Caplan, E.H. Poindexter, B.E. Deal, R.R. Pazouk: ESR centers, interface states, and oxide fixed charge in thermally oxidized silicon wafers. *J. Appl. Phys.* **50**, 5847 (1979)
36. E.H. Poindexter, G.J. Gerardi, M.E. Rueck, P.J. Caplan, N.M. Johnson, D.K. Biegelsen: Electronic traps P_b centers at the Si/SiO₂ interface: Band-gap energy distribution. *J. Appl. Phys.* **56**, 2844 (1984)
37. H. Boroffka, E.F. Krimmel, M. Lindner, H. Runge: The origin of leakage current of laser and electron beam annealed diodes. In: *Proc. Laser and Electron Beam Processing of Electronic Materials*, Proceeding Volume 80-1 (Electrochemical Society, Pennington, NJ 1980) pp. 178–186
38. N.M. Johnson, D.K. Biegelsen, M.D. Moyer: Low-temperature annealing and hydrogenation of defects at the Si–SiO₂ interface. *J. Vac. Sci. Technol.* **19**, 390 (1981)
39. E.F. Krimmel: Silicon nitride: electronic structure; electrical, magnetic, and optical properties; spectra; analysis. In: *Gmelin Handbook of Inorganic and Organometallic Chemistry, Si, Silicon*, Supplement Vol. B 5b2, 8th edn (Springer, Heidelberg 1997), pp. 25–52, 164–170
40. *Proc. First Intern. Conf. on Ion Implantation in Semiconductors*, Thousand Oaks, USA (1970); *Proc. Second Intern. Conf. on Ion Implantation in Semiconductors*, Garmisch Partenkirchen, Germany (Springer, Heidelberg 1971); *European Conf. on Ion Implantation* (Peter Pelegrinus, Reading 1970)
41. H. Müller, H. Ryssel, I. Ruge: A new method for boron doping of silicon by implantation of BF₂-molecules. In: *Proc. Second Intern. Conf. on Ion Implantation in Semiconductors*, ed. by I. Ruge, J. Graul, Garmisch-Partenkirchen, Germany (Springer, Berlin, Heidelberg 1971) pp. 85–95
42. H. Boroffka, E.F. Krimmel, H. Mader, H. Runge: Laser annealing of semiconductor devices. In: *Proceedings of Laser Effects in Ion Implanted Semiconductors* (Catania 1978) pp. 224–231
43. E.F. Krimmel: Slip line free silicon in large-area multiple-scan annealing with a line-focused electron beam. *Phys. Stat. Sol. (a)* **70**, K63 (1982); E.F. Krimmel: Elektroneninterferenzen in der Umgebung der Brennlinie einer magnetischen Quadrupollinse. *Z. Phys.* **163**, 339 (1961)
44. E.F. Krimmel, H. Oppolzer, H. Runge, W. Wondrak: Kinetics of scanned electron beam annealing of high-energy as ion implanted silicon. *Phys. Stat. Sol. (a)* **66**, 565 (1981)
45. E.F. Krimmel: Research on future microcircuit technology. *Trans. South African Inst. Elec. Eng.* **74**, 128 (1983)
46. E.F. Krimmel: Verfahren zur Oberflächenvergrößerung eines Substrates. Offenlegungsschrift DE 33 10 331 A1 (1984)
47. A. Markwitz, H. Baumann, E.F. Krimmel: Height control of silicon nano whiskers embedded in ultra thin silicon nitride layers by rapid thermal annealing. *Physica E* **11** 110 (2001)
48. H. Boroffka: privat communication
49. D. Bäuerle: *Laser Processing and Chemistry*, 3rd edn (Springer, Heidelberg 2000)
50. M. Winstel, E.F. Krimmel, A. Weiss: Excimer laser-induced deposition of silicon using SiHCl₃ precursor. *Siemens Forsch.- u. Entwickl.-Ber.* **17**, 6 (1988)

51. G. Dearnaley, J.H. Freeman, R.S. Nelson, J. Stephen: *Ion Implantation* (North Holland, Amsterdam 1973)
52. H. Ryssel, I. Ruge: *Ionenimplantation* (Teubner, Stuttgart 1978)
53. S.M. Sze: *Physics of Semiconductor Devices* (Wiley, New York 1981)

12 Neutron Transmutation Doping (NTD) of Silicon

M. Schnöller

12.1 Introduction

For the construction of high-power devices such as the high-power thyristors that are commonly used for the control of motor drives for engines and rolling mills and for high-power DC transmissions, large-sized silicon crystals having an extremely tight phosphorus background doping corresponding to about $40\Omega\text{ cm}$ or higher are required. The exact level of the phosphorus concentration and its homogeneous distribution are decisively responsible for the electrical quality of the thyristor: the breakdown behaviour and blocking voltage depend upon the maximum of the doping concentration, while the minimum of the doping concentration is responsible for the quality of the high-temperature behaviour of the device. Optimal properties can be expected if the distribution of the dopant is exactly homogeneous and no difference between the maximum and the minimum of the dopant concentration exists.

Moreover, the exact reproduction of doping is important for the construction of high-voltage equipment for high-energy plants, where a large number of devices with identical properties are required.

12.2 Conventional Phosphorus Doping

Conventional phosphorus-doping processes such as gas-phase doping show particular difficulties in the doping region of about $40\Omega\text{ cm}$ and above, concerning the exact dopant concentration, its homogeneous distribution and the exact reproducibility of the doping process. Measurements of the electrical resistivity also show that the phosphorus dopant is distributed inhomogeneously in the silicon crystal. Macroscopic and superimposed periodic microscopic oscillations of the electrical resistivity are observed in the radial and axial directions of the silicon crystal rods [1, 2].

Figures 12.1a and 12.2 left show the lateral macroscopic and microscopic inhomogeneous distribution of phosphorus in conventionally doped silicon on the basis of measurements of the electrical resistivity [3]. Measurements of the avalanche radiation of diodes performed by Voss [4] (Figs. 12.3b and 12.3d) demonstrate the lateral and axial macroscopic inhomogeneous distribution of

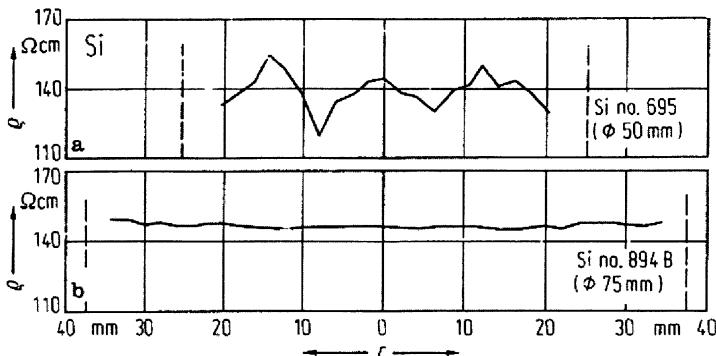


Fig. 12.1. Lateral microscopic distribution of the resistivity ρ versus radial distance r from the centre of a silicon wafer (four-point probe measurement). (a) Conventionally doped Si; (b) neutron-irradiated Si [3]

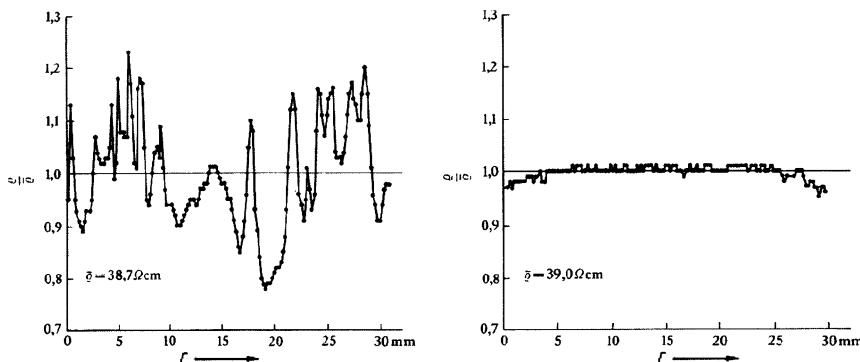


Fig. 12.2. Lateral microscopic variation of the relative resistivity $\rho / \bar{\rho}$ versus radial position r in a silicon wafer. $\bar{\rho}$ is the average resistivity. The measurement of the spreading resistance was performed with Al-Si contacts [18]. *Left* Conventionally doped Si; *Right* neutron-irradiated Si [3]

the dopant, in contrast to the homogeneous distribution of the phosphorus-dopant in neutron-irradiated silicon (Figs. 12.1b, 12.2 right, 12.3a and 12.3c).

Microscopic periodic oscillations concerning the distribution of built-in impurities are also observed in crystals of materials other than silicon grown by the method of zone melting with rotation. Witt and Gatos [5] showed evidence that the observed impurity striations are due to a periodic alternation of the growth velocity. Hurle et al. [6] found that alternations of the growth velocity cause alternations of the concentration of the built-in impurities. The rotation of the growing crystal in an inhomogeneous temperature field was finally identified by Morizane et al. [7] to be the cause of these oscillations.

It can be assumed that this effect also causes the micro-oscillations of the phosphorus concentration in silicon crystals.

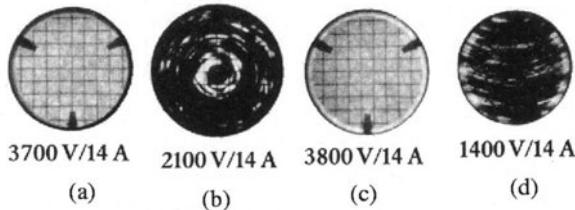


Fig. 12.3. Avalanche radiation from diodes made from neutron-irradiated Si (a), (c) and conventionally doped Si (b), (d). Horizontal section (a), (b); vertical section (c), (d). The homogeneous distribution of the electrical resistivity in diodes made from neutron-irradiated Si results in a homogeneous distribution of the avalanche radiation (*greyish appearance*). The inhomogeneous resistivity of the conventionally doped diodes shows local avalanche radiation (*black and white contrast*). The vertical section (d) shows the shape of the border between the liquid and solid phases in the growing Si crystal [4]

12.3 Phosphorus Doping by Means of Neutron Irradiation [8]

12.3.1 History

During our work to eliminate resistivity striations in conventionally phosphorus-doped silicon crystals, we performed annealing experiments slightly below the melting point of silicon. The silicon crystals were annealed for two weeks in an annealing furnace. After measuring the resistivity striations, we could not detect any significant difference in the distribution of the resistivity before and after the annealing. This poor result made us look for a doping method that would put in the phosphorus after the growth of the silicon crystal.

The idea of putting in the phosphorus by means of ion implantation was unsuitable because of the low implantation depth of the charged particles. We believed the neutron to be the only particle to have a chance of penetrating the lattice of a silicon crystal without any significant hindrance. Fortunately there is also a reaction of neutrons with silicon to generate phosphorus.

We felt it was an act of friendliness by nature that the values for the cross-section of the reaction, the half-life for the decay of the unstable intermediate and the strength of the side reactions were located in a region acceptable for an industrial technique, although it was expected that severe crystal defects could be generated during the irradiation, and therefore it was feared that the crystal would be irreparably destroyed. After we had started our investigations, we encountered the pioneering work of Lark-Horovitz [9] and Tanenbaum [10] in the literature.

Lark-Horovitz was the first to propose, in 1951, the generation of dopant by means of neutron irradiation and performed the first experiments with

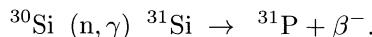
silicon crystals. But measurements of the electrical resistivity did not show the expected n-conductivity, because the crystal defects generated during the neutron bombardment had not been healed.

Tanenbaum and his collaborators manufactured diodes from neutron-irradiated silicon with resistivities in the region of $\sim 10 \Omega \text{ cm}$ in 1961. But at that time the required resistivity region could be served well with the aid of conventional doping methods and the requirements on the quality of the doping were modest at this time. Besides, there was no need for phosphorus-doped silicon crystals in higher-resistivity regions. It was also a disadvantage that Tanenbaum et al. did not wait for the unstable phosphorus isotope ^{32}P , generated in a significant amount in this doping region during irradiation, to decay, and the silicon therefore had some undesired radioactivity. This caused this method of doping soon to fall into oblivion.

12.3.2 Doping Reactions

The element silicon consists of the isotopes ^{28}Si (92.23%), ^{29}Si (4.67%) and ^{30}Si (3.10%). During irradiation with neutrons of thermal energy, the isotopes ^{28}Si and ^{29}Si form the non-radioactive isotopes ^{29}Si and ^{30}Si .

In contrast to these reactions, the silicon isotope ^{30}Si reacts with a neutron to form the unstable isotope ^{31}Si , which decays with a half-life of 2.6 hours to the stable phosphorus isotope ^{31}P :



Because of the low cross-section for the reaction ($\sigma_{abs} \sim 0.16 \text{ barn}$), the spatial distribution of the phosphorus generated in the silicon crystal is – for practical purposes – independent of the dimensions of the irradiated crystal. The dimensions of the crystal rods to be irradiated are limited only by the dimensions of the irradiation facility. The spatial distribution of the generated phosphorus is due to the homogeneity of the spatial neutron density distribution, and the neutron fluence defines the phosphorus concentration. All these conditions are easily adjustable. Therefore this doping method is very appropriate for industrial use, particularly concerning resistivity regions of about $40 \Omega \text{ cm}$ and higher, where conventional methods have increasing difficulties in the accuracy of attaining the precise doping level, and in achieving homogeneity of the dopant distribution and reproducibility of the doping process.

Because of the high neutron doses required to generate the required phosphorus concentration, the irradiations are commonly performed in research reactors with a high neutron flux density, preferably in heavy-water-moderated facilities. It takes about one hour of irradiation time to attain a doping level corresponding to a resistivity of $100 \Omega \text{ cm}$. Figure 12.4 shows the relations between the time of irradiation, the concentration of phosphorus generated and the electrical resistivity required, using a neutron flux density of $8 \times 10^{13} \text{ cm}^{-2} \text{ s}^{-1}$.

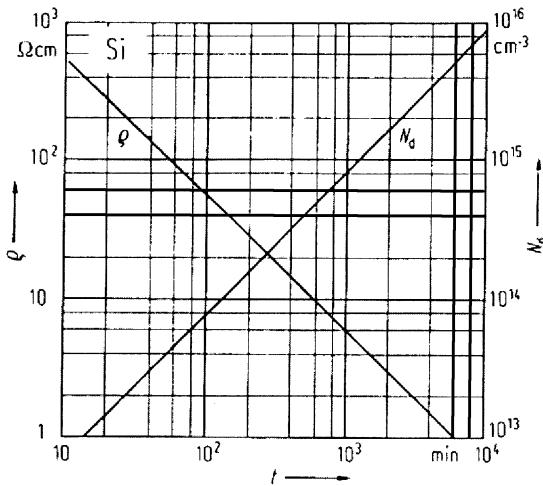
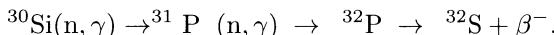


Fig. 12.4. Calculated relation of the irradiation time t to the resistivity ρ (left axis) and to the number of phosphorus atoms N_d generated (right axis) after an irradiation with a thermal-neutron flux density of $8 \times 10^{13} \text{ cm}^{-2} \text{ s}^{-1}$; calculated with the help of the Irvin curve [11, 19]

12.3.3 Side Reactions

The generation of the unstable phosphorus isotope ^{32}P , which is formed by a secondary reaction of ^{31}P with a neutron, has to be taken into account as an important side reaction during neutron irradiation. This isotope decays with a half-life of 14 days to form the stable sulphur isotope ^{32}S :



The concentration of the generated phosphorus isotope ^{32}P depends not only upon the time of irradiation and the cross-section for the reaction, but also upon the neutron flux density in the reactor during irradiation. Figure 12.5 [11] shows the relations between the neutron flux density during irradiation, the desired electrical resistivity and the concentration of the generated phosphorus isotope ^{32}P .

Apart from the reactions of thermal neutrons with silicon, there are a few reactions with the fast neutrons from the reactor. Fortunately these reactions have low cross-sections and the radioactive intermediates generated have short half-lives. The low concentrations of the stable final products generated magnesium and aluminium, have no practical importance for the phosphorus doping of silicon. If, however, there are impurities in the silicon crystals to be irradiated, for example the element indium, which is used in the production of infrared detectors, the reactions of these elements with neutrons have to be taken into account.

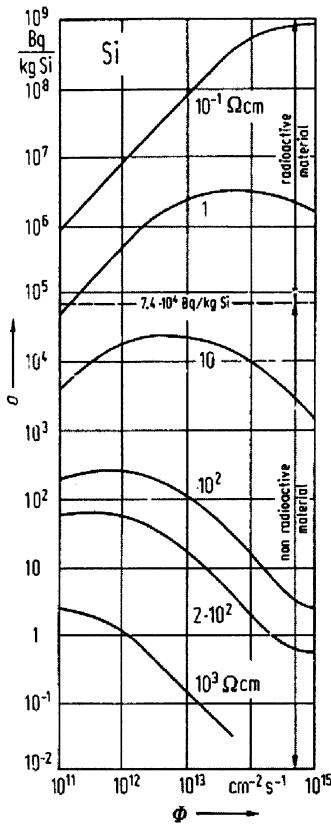


Fig. 12.5. Relationship between the neutron flux density Φ and ^{32}P radioactivity a for various resistivities [11]

12.3.4 Radioactivity of the Irradiated Silicon

In general, for the doping region of interest in industrial use, there are no significant problems concerning the radioactivity of the irradiated silicon. Because of the short half-life of the unstable intermediate ^{31}Si , the irradiated silicon can usually be handled like unirradiated material after a decay time of 3 days.

In the region of $5\ \Omega\text{cm}$ and below, which is generally less interesting for industrial doping processes, longer decay times arising from the unstable phosphorus isotope ^{32}P have to be taken into account. The relations between the required resistivity, the mass of irradiated silicon and the decay time are presented in Fig. 12.6 [12].

If there are radioactive impurities, caused for example by contaminated water in the reactor, they can be etched away easily with a mixture of

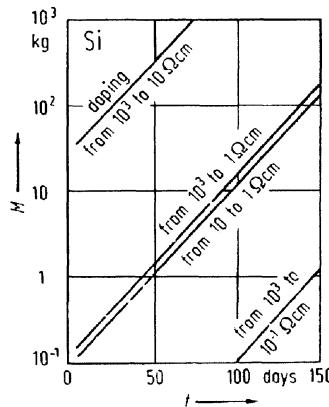


Fig. 12.6. Decay time t of ^{32}P in neutron-irradiated Si for reaching the acceptable limit ($a \sim 3.7 \times 10^5 \text{ Bq}$) versus the quantity of irradiated Si M for various resistivities. (The calculation was performed for a neutron flux density of $\Phi = 10^{14} \text{ cm}^{-2} \text{ s}^{-1}$) [12]

HNO_3/HF . Normally the surface of the irradiated silicon is completely free from contamination products.

12.3.5 Annealing of Crystal Defects

During irradiation, various defects in the silicon crystal are generated, mostly caused by fast neutrons. The electrical resistivity of the crystals is about $10^5 \Omega \text{ cm}$. In practice, the crystal defects can be healed completely by annealing at higher temperatures. Concerning the minority carrier lifetimes, there exists no difference between conventionally doped and annealed neutron-bombarded silicon [13]. The lower the content of fast neutrons during irradiation, the lower is the annealing temperature necessary to heal the crystal defects. Typical annealing conditions are a temperature of about 750°C and 1 hour annealing time [14]. As an example of the annealing behaviour, Fig. 12.7 [15] presents the development of the phosphorus donor lines in the electronic excitation spectra for different annealing temperatures. At a temperature of 800° , the regeneration of the crystal lattice is complete. It is also possible to heal the crystal defects with the help of infrared light [16]. Further details of crystal defects and their healing are reported in [8].

12.3.6 Technological Implementation of Silicon Doping

Nuclear Reactors

Because of the high neutron flux density necessary to generate the required phosphorus concentration, the irradiations are mostly performed in

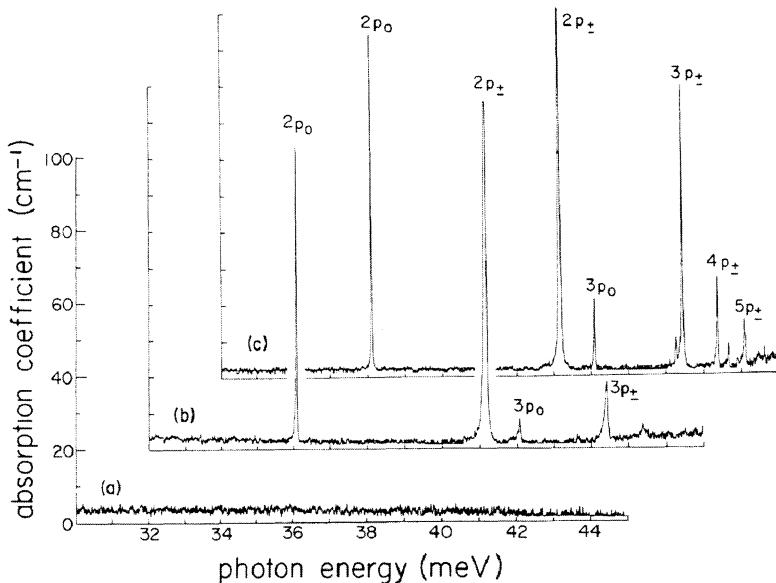


Fig. 12.7. Annealing of defects produced during the NTD process. The concentration of phosphorus generated was $\sim 2 \times 10^{15} \text{ cm}^{-3}$. (a) Before annealing. (b) After annealing for 2 h at 650°C . (c) After annealing for 1 h at 800°C [15]

Table 12.1. Nuclear reactors used for the generation of NTD silicon, and their irradiation capabilities [17]

(HW = heavy-water-moderated reactor, LW = light-water-moderated reactor)

Saclay, France	HW	30 t/a
Studsvik, Sweden	LW	30 t/a
Risø, Denmark (shut down)	HW	40 t/a
Kjeller, Norway	HW	12 t/a
Mol, Belgium	LW	20 t/a
Missouri, USA	LW	15 t/a
MIT, USA	HW	10 t/a
Lucas Heights, Australia	HW	18 t/a
Other reactors	HW/LW	~ 10 t/a
Sum	LW	70 t/a
Sum	HW	75 t/a

nuclear reactors operated by research institutes or universities. Heavy-water-moderated reactors are preferred, because of their high yield of neutrons with low energies. Irradiation usually takes a couple of hours. In Table 12.1 [17], reactors which have been used are listed.

The new high-flux heavy-water-moderated reactor FRM II at the Technical University of Munich operating in 2004 provides irradiation facilities for silicon rods with a length of 500 mm and a diameter of 200 mm. The neutron flux density is $2 \times 10^{13} \text{ cm}^{-2} \text{ s}^{-1}$ and the spatial homogeneity is about $\pm 5\%$. The ratio of low-energy neutrons to high-energy neutrons amounts about 1000:1, and a capacity of 10^4 kg per year is possible. First commercial irradiations for silicon doping are planned for the year 2005.

Amounts of Irradiated Silicon

From the beginning of the irradiation of silicon for commercial use in 1973 [20] to the present date, the quantities of silicon irradiated have increased from the region of some kilograms [11] to the region of some tens of tons per year [17]. The main reason is that silicon has been irradiated not only to achieve resistivities in the region of $100 \Omega \text{ cm}$ and above, where conventional doping methods begin to fail, but also to achieve lower resistivities, because of the excellent advantages of the irradiated material. The higher costs of the doping were overcompensated by higher yields for obtaining the target doping and by improved efficiencies of the manufactured devices within tight limits.

Figure 12.8 [17] shows the quantities of silicon irradiated over a period of about 20 years. The decreases in the quantity for some years were caused not

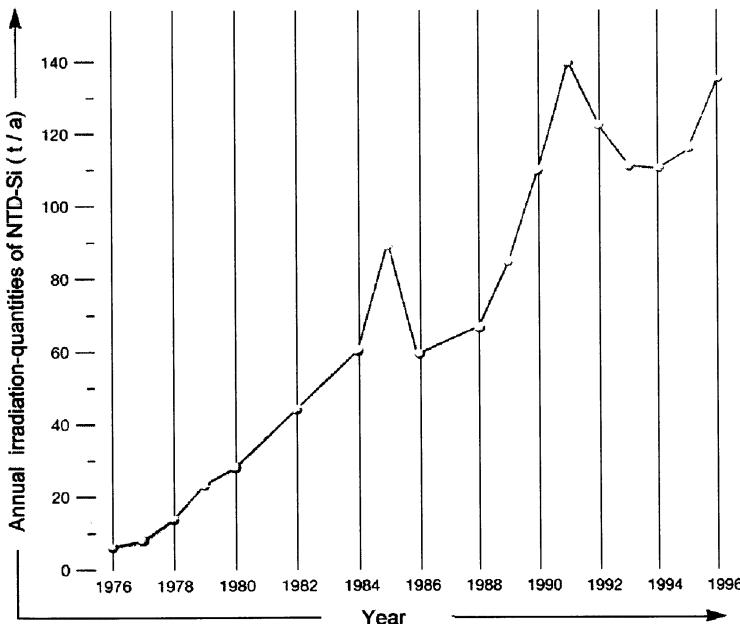


Fig. 12.8. Annual quantities of neutron-irradiated silicon [17]

only by commercial demand for silicon devices but also by improvements in the more economical technology of the gas-phase doping process in achieving the aims of doping in the region of higher resistivity values and of improved homogeneity in the distribution of the phosphorus.

In past years there have frequently been problems caused by the decommissioning of irradiation facilities. These problems could be solved up to now by increasing the capacities of the remaining reactors. In the future it has to be taken into account that an increasing number of irradiation facilities will have to be closed because they have reached their limits of age, such as the Risö reactor, which has already been shut down prematurely. This is particularly true for heavy-water-moderated reactors, which are preferentially used for commercial doping. Therefore a lack of irradiation capability is to be expected in the near future.

12.4 A Forward Look

Even in the future, there will be a need for silicon crystals doped with the aid of neutron irradiation, especially in the region of higher electrical resistivities. It may be expected that the commercial boundary between the favourable use of conventional gas-phase doping and of neutron-irradiation doping may be shifted towards the cheaper gas-phase doping method, because of technical advances in the conventional method. At present it can be estimated that the region of higher electrical resistivities will be dominated by NTD silicon in future as well. It can also be expected that the requirements concerning the accuracy, homogeneity and reproducibility of doping will increase as a result of the development of future silicon devices. This means that NTD Silicon will be used not only in resistivity regions accessible exclusively by neutron-doped silicon, but also once again in regions which are well served by the conventional method.

References

1. J. Burtscher: Resistivity fluctuations, growth striations and swirls in silicon single crystals. Proceedings of the European Summer School, July 8–12, 1974, Bad Boll, ed. by H. Weiss (1974) pp. 63–91
2. E. Spenke, W. Heywang: Phys. Stat. Sol. (a) **64(11)**, 12 (1981)
3. M. Schnöller, E. Spenke: Physik in unserer Zeit **7(1)**, 1 (1976)
4. P. Voss: IEEE Trans. Electron. Devices **ED-20**, 299 (1973)
5. A.F. Witt, H.C. Gatos: J. Electrochem. Soc. **115**, 70 (1968)
6. D.T.J. Hurle, E. Jakeman, E.R. Pike: J. Cryst. Growth **3/4**, 633 (1968)
7. K. Morizane, A.F. Witt, H.C. Gatos: J. Electrochem. Soc. **114**, 738 (1976)
8. M. Schnöller: Nuclear transmutation doping. In: *Landolt-Börnstein*, New Series Vol. 17, *Semiconductors*, Subvolume c, ed. by M Schulz, H Weiss (Springer, Berlin, Heidelberg 1984) Sect. 6.1.4.3., pp. 185–191 and 513–516

9. K. Lark-Horovitz: Nucleon-bombarded semiconductors. In: *Semiconducting Materials, Proceedings of a Conference at Univiversity of Reading* (Butterworth, London 1951) pp. 47–69
10. M. Tanenbaum, A.D. Mills: J.Electrochem. Soc. **108**, 171 (1961)
11. E. Haas, J.A. Martin: Nuclear transmutation doping from the viewpoint of radioactivity. In: *Neutron Transmutation Doping in Semiconductors*, ed. by J.M. Meese (Plenum Press, New York and London 1979) pp. 27–36
12. E.W. Haas, M.S. Schnöller: J. Electron. Mater. **5**, 57–68 (1976)
13. P. Voss: private communication (1976)
14. J. Corish, F. Benière, V.K. Agrawal, S. Harridos, C. Defeux: J. Appl. Phys. **50**, 6338 (1979)
15. C. Jagannath, Z.W. Grabowsky, A.K. Ramdas: Phys. Rev. B **23**, 2082 (1981)
16. A. Yusa, D. Itoh, C. Kim, H. Kim, K. Hushimi, S. Ohkawa: Application of NTD silicon for radiation of surface barrier type. In: *Neutron Transmutation-Doped Silicon. Proceedings of the Third International Conference on Neutron Tansmututation Doping of Silicon*, Aug. 27–29, 1980, Copenhagen, ed. by J. Guldberg (Plenum Press, New York 1980) pp. 473–485
17. W.V. Ammon: *Neutronen in der Silizium-Haibleitertechnik, Neue Forschungsneutronenquelle Garching* (Technische Universität München, Garching 1998) pp. 78–80
18. J. Krauß: National Bureau of Standards Special Publication 400-10, *Spreading Resistance Symposium*, Proceedings of a Symposium held at NBS, Gaithersburg, MD, June 13–14 (1974)
19. J.C. Irvin: Bell Syst. Tech. J. **41**, 387 (1962)
20. M. Schnöller: IEEE Trans. Electron Devices **21**, 313 (1973)

Part VI

The Roles of Certain Impurities

13 Transition Metal Impurities in Silicon

T. Heiser

13.1 Introduction

Unlike dopants, metal impurities are rarely used in semiconductor devices. Nevertheless, transition metals are technologically important, since their presence in silicon is difficult to avoid and their interaction with free charge carriers may influence electronic devices [1]. An illustrative example is given by copper in silicon. As few as 10^{12} copper atoms per cubic centimeter of silicon may lead to significant yield loss in submicron integrated circuit manufacturing [2–4]. The high electrical and thermal conductivity of copper thin films has nevertheless pushed the electronics industry to introduce copper interconnects into its products in spite of the enhanced risk of copper contamination. Only a perfect control of copper impurities in silicon technology allows the industry such a performance. Another technological issue related to transition metals appears in the silicon photovoltaic industry, where cost reductions are pursued by using low-quality silicon with large amounts of metal impurities [5, 6]. Defect reactions occurring during device processing determine the final energy conversion efficiency. Understanding the reaction paths of the dominant impurities, their electrical activity in various states and their mutual interaction has helped in optimizing the performance of photovoltaic devices.

Today, the concentration of metal impurities in as-grown high-quality silicon wafers designed for ULSI (ultra large-scale integration) technology generally does not exceed 10^{11} at/cm³. Contamination occurs mostly during processing steps of silicon wafers such as high-temperature treatments, ion implantation, chemical cleaning and wafer handling. Fe, Cu and Ni are the most frequently observed contaminants. The potential for contamination during silicon processing depends on physical properties such as the electronegativity, vapor pressure, bulk diffusion and solubility in silicon of the impurity. The larger electronegativity of Cu (1.9) compared with Si (1.8), for instance, triggers surface deposition in the liquid phase, making chemical processing the dominating pathway for copper contamination. Large diffusion coefficients and solubilities (at the process temperatures) allow adsorbed surface atoms to diffuse rapidly into the silicon and escape from subsequent surface cleaning. In contrast, a high vapor pressure favors surface evaporation and reduces bulk contamination. This is the case for Zn and Hg, as well as for several oxidized metal species [7].

The consequences of contamination for the final device depend on the spatial distribution of the contaminants. Metal impurities may either accumulate at interfaces, be included into silicon oxide or end up in the bulk. In the latter case, the proximity to the active region of a device is a determining factor. For most electronic devices, only the first few microns below the surface are of importance and need to be kept clean. Any metal contaminants located deeper in the material will have no effect on device operation. In photovoltaic devices, however, the whole wafer participates in light absorption and charge transport, and needs to be shielded. The interaction between metal impurities and free charge carriers depends on the microscopic state of the impurity. Dissolved and precipitated impurities behave differently. Isolated impurities introduce localized electronic states, which may either trap minority carriers or enhance their recombination rate. This can lead to low current gains in bipolar transistors and to low conversion efficiencies in photovoltaic devices. On the other hand, metal precipitates are mostly responsible for low-resistance paths and may yield high leakage currents and low breakdown voltages in rectifying devices. When located in the near-surface region, metal precipitates may also cause local oxide thinning and induce premature voltage breakdown.

The current understanding of the physical properties and electrical activity of metal impurities in silicon and the engineering solutions which are currently used by the industry to control their influence on electronic devices are the outcome of numerous research efforts performed worldwide over the last 50 years. A number of review articles and books on metal impurities in silicon have been published over this period of time (see for instance [8–10]) and have helped semiconductor scientists and engineers to keep up with the rapid evolution of the experimental and theoretical results. The present chapter is not designed to be another review article, but to give the noninitiated reader a comprehensive introduction to the physics of transition metal impurities in silicon. Fundamental concepts are illustrated by well-established results, while references to the most recent publications in the field are added to allow further investigation by the interested reader.

The chapter is organized as follows. In the first section, the diffusion mechanisms and equilibrium solubility of the major transition metals are addressed. The second section describes the electrical activity of metal impurities and associated device failures, while the last section deals with state-of-the-art impurity engineering, such as impurity gettering and metal trace detection.

13.2 Diffusion and Solubility

The diffusion coefficient and solubility of a metal impurity in silicon are its most consequential properties. They determine the impurity penetration depth and bulk concentration which may result from accidental surface contamination and subsequent high-temperature annealing. Fundamental studies

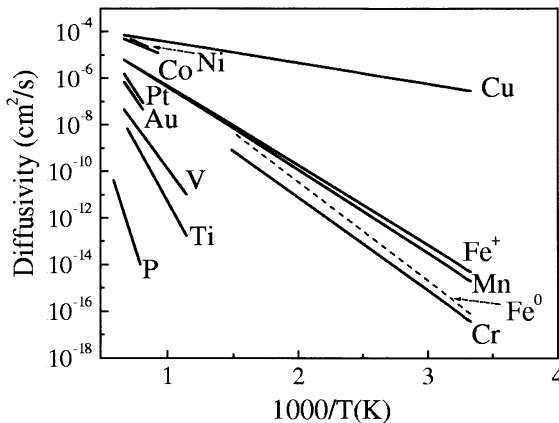


Fig. 13.1. Diffusion coefficient versus reciprocal temperature for various impurities (from [7], except for the Cu data, which are from [15])

of these properties for major contaminants were done in the 1960s [8, 11–13]. Most of these data are still valid today, with the exception of copper in silicon, whose diffusion coefficient had been considerably underestimated [14, 15]. Figure 13.1 represents up-to-date values for the diffusion coefficients of major metallic impurities as a function of temperature. The data for phosphorus impurities have been added for comparison.

The significantly larger diffusion coefficients of transition metals as compared with phosphorus point to different diffusion mechanisms. In thermal equilibrium, shallow dopants are known to occupy exclusively substitutional positions. This energetically stable configuration allows dopant diffusion steps to occur only when an intrinsic point defect (self-interstitial or vacancy) is localized at a nearest-neighbor position. The point defect equilibrium concentration drops quickly with temperature and is responsible for the extremely slow dopant diffusion below 900°C. In contrast, the fast diffusion of transition metals suggests that, at the diffusion temperature, a significant amount of the impurity is located on interstitial sites and diffuses rapidly through the open crystal structure of silicon. The diffusion coefficients shown in Fig. 13.1 are functions of the interstitial-to-substitutional ratio of the metal (substitutional impurities being considered as immobile) and of the diffusivity of the interstitial component.

It is currently accepted that 3d transition metals have dominant interstitial solubilities at all temperatures [7]. The chemical trend followed by the migration enthalpy (the slope of the lines in Fig. 13.1), which decreases with increasing atomic number, was attributed by Utzig to the influence of the elastic energy of the impurity in the silicon host lattice [16].

The numerous data published on Fe in silicon (see [17] for a recent review) led to the general agreement that iron has an essentially interstitial

equilibrium solubility and diffuses interstitially with a migration enthalpy of 0.68 eV. Despite the rather scarce experimental results concerning Ni and Co (mainly owing to the difficulty of keeping these impurities in the interstitial sites at room temperature), it is also believed that these impurities dissolve and diffuse interstitially at high temperatures [8].

For copper, Hall and Racette [11] found that in p-type and in moderately doped n-type material more than 99% of the metal remains on interstitial sites under thermal equilibrium conditions, and estimated the interstitial copper migration enthalpy as 0.43 eV. However, recent experiments showed that this value was a significant overestimate [18] owing to copper–acceptor pairing (see below). An intrinsic copper migration enthalpy of only 0.18 eV was established recently by Istratov et al. [14, 15] using drift experiments in low-doped material. This value agrees well with ab-initio calculations done by Woon et al. [19]. In high-doped material or at low temperatures, the copper–acceptor interaction must be taken into account. The *effective* diffusion coefficient calculated assuming a diffusion-limited pairing reaction and a thermally activated dissociation rate agrees well with Hall and Racette's data [15].

The diffusion profiles observed in samples contaminated with Au or Pt could not be understood in terms of a simple interstitial diffusion mechanism. Gösele et al. [20] found that these metals do have a significant substitutional equilibrium solubility and need to be “kicked out” into interstitial sites by interaction with silicon self-interstitials before being able to diffuse. This kick-out mechanism is described by the following reaction:



where M_s and M_i correspond to the fixed substitutional and mobile interstitial metal impurity, respectively, while I represents a silicon self-interstitial. This reaction results in a nonuniform *effective* diffusion coefficient D_M^{eff} , defined by

$$D_M^{eff} = D_{M_i} f_i \quad (13.2)$$

where $f_i = [M_i]/([M_i] + [M_s])$. The terms in brackets represent the impurity concentration on different atomic positions. Equation (13.2) emphasizes that only the interstitial impurity is mobile. The factor f_i depends *locally* on reaction (13.1). Under nonequilibrium conditions, self-interstitial supersaturation tends to increase f_i , and undersaturation tends to decrease it. In the particular case where the point defect concentration is pinned to its equilibrium value (through a high dislocation density, for instance), f_i is uniform and leads to the constant (concentration-independent) effective diffusion coefficients shown in Fig. 13.1 [21, 22]. Experimental values for the *intrinsic* interstitial diffusion coefficient (D_{M_i}) of these substitutional transition metals have not been reported so far.

The contribution of point defects to the diffusion mechanism couples the diffusion of a substitutional metal to the local defect concentration, which itself may be affected by technologically important processes such as thermal

oxidation or ion implantation or by the presence of extended defects (interfaces, dislocations, grain boundaries, . . .). For instance, self-interstitial injection during oxidation increases the interstitial metal concentration through reaction (13.1) and enhances the diffusion of the metal. On the other hand, in the neighborhood of extended defects the lower point defect concentration reduces the effective diffusion coefficient and favors metal accumulation. Gettering of substitutional metals, discussed later in this chapter, partly relies on this point-defect–impurity interaction.

In semiconductors, defects can have various charge states, depending on the Fermi level position and on the defect ionization energy. Accordingly, the charge state may influence the impurity diffusivity either through a charge-state-dependent D_{M_i} , as has been found in the case of Fe [23, 24], or by a Fermi-level-dependent interstitial solubility. An example illustrating the latter effect is given by copper in silicon. Hall and Racette [11] showed that interstitial copper impurities behave as shallow donors, being positively charged at room temperature independently of the Fermi level position. Conversely, substitutional copper impurities, although a minority, were found to exist mostly in a negatively charged state (an acceptor-like defect). Upon introduction into the silicon matrix, ionization of copper atoms induces a Fermi-level-dependent energy gain. In p-type material, electron release from interstitial Cu is the energetically favorable path and increases the *interstitial* solubility. In n-type doped material, however, ionization of the substitutional acceptor levels, when located below the Fermi level, becomes energetically favorable. This increases the *substitutional* solubility, which even exceeds the interstitial solubility in highly doped n-type material. In this case, diffusion of copper is significantly slowed down and is most probably coupled to the intrinsic point defects, as is the case for other substitutional metals such as Au or Pt.

The solubility of a given impurity in silicon is defined as the impurity concentration dissolved in the silicon matrix in equilibrium with a second phase, in most cases a metal silicide [8]. For instance, Fe and Ni were found to be in equilibrium with FeSi_2 and NiSi_2 over the whole temperature range investigated, while the composition of the copper boundary silicide is Cu_3Si [25–27]. All transition metals are characterized by a solubility which decreases rapidly with temperature, reaching negligible values at room temperature [8]. Dissolved metal impurities in silicon observed at room temperature thus always constitute a metastable supersaturated solution. The level of supersaturation depends on various factors such as the cooling rate from the in-diffusion temperature, the impurity diffusion rate, the crystallographic structure of the silicide and the presence of secondary defects. Interesting examples are given by Fe and Ni, whose behaviors during cooling differ significantly. The lattice parameters of Ni silicide are close to those of silicon, leading to low elastic strain during the formation of silicide particles [26]. This in turn favors rapid precipitation of supersaturated metal and is thought to keep the interstitial nickel concentration below the detection limit at room temper-

ature even after extremely fast cooling (2000°C/s). In contrast, the lower diffusion rate of Fe and the larger volume change associated with formation of the silicide FeSi_2 as compared with the Ni silicide allows significant metal supersaturation after cooling [25]. Nucleation of Fe silicide precipitates has been observed to be triggered mostly by secondary defects (dislocations, interfaces, ...), which reduce the Si/silicide interfacial energy through strain relaxation. Subsequent precipitate growth is limited, however, by long-range impurity diffusion, which constitutes a stronger limitation for Fe than for Ni. Consequently, moderately fast cooling rates ($< \sim 50^{\circ}\text{C/s}$) are generally enough to avoid the formation of iron precipitates in high-quality crystals.

The behavior of Cu is different again. Initially, Cu precipitation was thought to occur almost quantitatively even during fast cooling, in spite of the significant lattice mismatch and the strong self-interstitial emission which necessarily accompanies the formation of metal-rich silicide particles [7]. The extremely high diffusion coefficient of interstitial copper was thought to be responsible for this behavior. Recently, however, the observation of interstitial copper through transient ion drift measurements and precipitation studies using synchrotron X-ray fluorescence led to the conclusion that thermal quenching at moderate cooling rates (larger than 50°C/s) can lead to a significantly supersaturated copper solution [28, 29]. Coulomb repulsion between the growing silicide particles and interstitial copper is at present believed to be responsible for this unexpectedly low precipitation rate. In the case of a slow cooling rate, copper precipitation occurs predominantly near the sample surface, which acts as a perfect sink for the emitted self-interstitials, and this precipitation can be visualized as haze after preferential surface etching [7].

Supersaturated dissolved metal impurities may also form complexes with other impurities, of identical or different chemical nature. Complex formation is expected if the cooling rate is too fast to allow quantitative precipitation and if the room temperature mobility of the dissolved metal impurities is large enough for them to be captured by secondary defects. Long-range electrostatic interactions also contribute to the complex formation rate. A typical example is given by the formation of donor-acceptor pairs [30]. In this case the opposite charge states of the two constituents give rise to a long-range Coulomb interaction, which allows the formation of stable defect pairs through a diffusion-limited trapping mechanism. It is well established, for instance, that at room temperature the interstitial diffusion coefficient of iron is large enough to let all Fe atoms pair with an acceptor dopant (B, Al, Ga, ...) within hours. Chantre and Kimerling [31] showed that the statistical distribution of these pairs among the various spatial configurations agrees well with a Boltzmann distribution assuming a point charge electrical potential. Similar pairs have been observed in Cr- and Mn-contaminated p-type silicon [32, 33]. Donor-acceptor pairs are generally weakly bound and can be dissociated upon annealing at relatively low temperature ($\sim 200^{\circ}\text{C}$). Since defect pairs and clusters have a different electrical activity from individual

metal impurities, they need to be taken into account in order to understand fully the influence of metal contamination on electronic devices.

The short description above is enough to give an idea of how the behavior of metal impurities may be coupled to typical device-processing steps. The solubilities and diffusion coefficients of most transition metals are high enough to allow metal impurities to in-diffuse from a contaminated surface over macroscopic distances (ranging from a few microns to the whole wafer thickness) during a typical high-temperature process. The final distribution and configuration of the metal impurities depend strongly on the cooling rate, which controls the level of supersaturation and triggers more or less the formation of metal precipitates. A strong coupling to intrinsic point defects, whose concentrations vary with the process parameters, is expected either through the impurity diffusion mechanism (this is the case for substitutional metal impurities such as Au or Pt) or through the precipitation kinetics. After a given process, metal impurities may either be in an interstitial (Fe, Cu, Ti, ...) or substitutional (Au, Pt, Cu in n^+ Si) state, be contained in a complex or be included in silicide particles. The tendency of some impurities to precipitate in the near-surface region, i.e. in the active region of devices, may be particularly detrimental to device operation.

13.3 Electrical Activity

The electrical activity of dissolved metal impurities is mostly due to localized energy states (or deep levels), which they introduce into the silicon energy bandgap. The interaction between these deep levels and the free charge carriers has been described by Shockley, Read and Hall in terms of carrier capture and emission rates (see Fig. 13.2) [34, 35]. The capture rates are given by

$$c_n = \sigma_n v_n n \quad \text{and} \quad c_p = \sigma_p v_p p, \quad (13.3a)$$

where $v_{n(p)}$ are the electron and hole thermal velocities, $\sigma_{n(p)}$ are the capture cross sections, and n (p) are the Fermi-level-dependent electron and hole concentrations, respectively. The deep-level emission rates are given by

$$\begin{aligned} e_n &= v_n \frac{N_c}{g} \sigma_n e^{-\Delta H_n/kT} \quad \text{and} \\ e_p &= v_p \frac{N_v}{g} \sigma_p e^{-\Delta H_p/kT}, \end{aligned} \quad (13.3b)$$

where N_c and N_v are the effective densities of states of the conduction and valence band, respectively, g is the deep-level degeneracy factor, and $\Delta H_{n(p)}$ are the deep-level electron and hole ionization enthalpies (the distance between the deep level and the conduction or valence band edge, respectively). Deep levels located close to the middle of the bandgap ($\Delta H_n \sim \Delta H_p$) allow transitions to and from both the valence and the conduction energy band

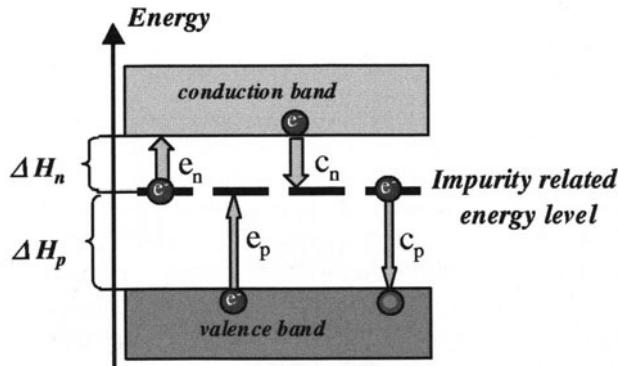


Fig. 13.2. Illustration of the interactions between a deep level and the silicon energy bands

equally, and promote the thermal generation of electron–hole pairs. This is particularly important if the impurity is located in a depletion region where carrier capture rates are negligible ($n, p \sim 0$). Such “generation centers” increase significantly the leakage current of PN junctions and Schottky barriers and reduce the charge storage time of charge-coupled devices. Similarly, deep levels with large capture rates (compared with the emission rates) for both electrons and holes lead to efficient recombination of electron–hole pairs. Since capture rates depend on the free-carrier concentration, recombination in doped materials occurs mostly under nonequilibrium conditions when the minority carrier concentration is increased either by photon absorption or by injection from a neighboring interface. Such “recombination centers” reduce the “lifetime” of excess minority carriers and lead to a low collection efficiency of photovoltaic devices. If located in the base of a bipolar transistor, recombinative impurities increase the base current and reduce the current gain of the transistor.

The capture cross sections may have values ranging from 10^{-18} to 10^{-14} cm^2 . This extended range reflects the influence of the impurity charge state on the capture process through the long-range Coulomb interaction. Owing to the Coulomb repulsion, for instance, electron capture by a negatively charged impurity is less likely than by a neutral or positively charged defect. Large electron capture cross sections thus attest to the donor nature of an impurity under investigation. The order of magnitude of the capture cross sections is often used to assess the acceptor- or donor-type character of impurities (see Table 13.1). Another, more explicit indication of the impurity charge state comes from the electric-field enhancement of carrier emission. This Poole–Frenkel effect is due to the distortion of the potential-energy distribution of an electron or hole around the impurity by the macroscopic electric field (in the depletion region of a Schottky barrier, for instance) [36]. The electric field lowers the energy barrier for electron emission from a donor-type defect or hole emission from an acceptor-type defect.

Table 13.1. Energy levels of major metallic impurities (from [7]). The subscripts i and s indicate the atomic position (interstitial or substitutional). The letter “d” or “a” indicates the electron donor or acceptor behavior of the deep level (“dd” corresponds to a double donor level and “aa” to double acceptor level)

Metal	Ti _i	V _i	Cr _i	CrB	Mn _i	MnB	Fe _i	FeB	Co _s	Ni _s	Cu _s	Cu _i – Cu _s	Au _s	Pt _s
ΔH_n (eV)	0.08 a 0.27 d	0.18 a 0.45 d	0.22 d		0.12 a 0.42 d	0.53 d	0.39 d		0.41 d	0.41 d	0.16 d		0.55 d	0.23 d
ΔH_p (eV)	0.28 dd	0.32 dd		0.29 d	0.27 dd			0.1 d	0.41 a	0.17 a	0.46 a 0.22 d	0.1 d	0.34 a	0.32 a

The currently most well-established data on the energy levels introduced by major metallic impurities are summarized in Table 13.1. Most metal impurities introduce recombination centers and must be kept at a low concentration when high minority carrier lifetimes and diffusion lengths are required. The recombination activity of Au has been put to use in high-speed switching devices, for which the recombination of excess minority carriers is the time-limiting process [37]. Interstitial copper is believed to introduce a relatively shallow donor level, whose precise location in the bandgap remains uncertain [9, 38]. From drift experiments, Prescha et al. found that in p-type silicon interstitial copper remains positively charged down to at least -70°C [18]. Substitutional copper is at present believed to introduce three deep levels corresponding to four different charge states [14]. The numerous other deep levels which have been observed in copper-contaminated silicon are due to defects complexes which include one or several copper atoms [39]. The formation of such defects is favored by the negligible room temperature solubility of copper in silicon, which constitutes a thermodynamic driving force, and by its fast diffusion, which lowers the kinetic barrier. Knowledge about the structure and composition of these complexes, however, remains rather scarce. The shallow donor level located at 0.1 eV above the valence band constitutes an exception. By observing the quadratic dependence of the deep-level concentration on the copper contamination level, Weber et al. [40] came to the conclusion that the corresponding defect is formed by an interstitial-copper–substitutional-copper pair. More recently, copper has been found to form complexes with other impurities such as H, Au and acceptor-like dopants [41–43].

Many efforts have been made to study the electrical activity of the technologically important impurity iron in silicon and have made this impurity one of the most well-understood defects in silicon. Interstitial iron is known to introduce a donor level at 0.39 eV above the valence band, which is responsible for the long-range Coulomb interactions between Fe and negatively charged defects in p-type material [7]. In boron-doped silicon, most Fe is paired with B and introduces a shallow level at 0.1 eV above the valence band [7]. Since the capture and emission rates of both interstitial iron and iron–boron pairs are known, changes in the minority carrier lifetime or diffu-

sion length subsequent to pair dissociation can be used to study the reaction kinetics quantitatively. Zoth and Bergholz suggested the use of this property to detect traces of iron in p-type silicon [44]. Thermal dissociation of these pairs is possible at temperatures above 200°C or at room temperature under high minority carrier injection by a recombination-enhanced reaction.

Since nickel precipitates almost quantitatively during cooling, its electrical activity is mostly due to nickel silicide particles, Ni-related defect clusters and to substitutional Ni. The electrical activity of nanometer-sized metal silicide particles has recently been investigated by Hedemann and Schröter [45]. These authors found that the temperature dependence of the emission rate can be understood in terms of band-like states located in the silicon bandgap. Their theory successfully reproduces the deep-level transient spectroscopy signature of both Ni and Cu precipitates. These band-like states can be considered as responsible for the recombination activity of precipitate particles and for the contrast in beam-induced current images. For larger particles, the bulk electrical properties of metal silicides dominate. Thus, a micron-sized copper or nickel silicide particle in silicon constitutes a metal/semiconductor interface, which may give rise to charge redistribution and related internal electric fields [46]. The low resistivity of metal silicides, further, makes them a preferential pathway for electric currents. If located in the active region of a device (most frequently at an Si/SiO₂ interface or in a space charge region), precipitates of fast-diffusing transition metals can be responsible for device failures either through recombination currents (from band-like states of small precipitates) or electrical breakdown [47].

13.4 Impurity Engineering

13.4.1 Gettering

Since metal impurities are electrically active and almost unavoidable, semiconductor device manufacturers try to minimize their impact on device operation by keeping them out of the active region. This can be achieved by introducing efficient impurity trapping sites either on the back surface of the silicon (for photovoltaic devices for instance) or just beyond the near-surface layer a few microns thick (for microelectronic devices). After contamination, the metal impurities diffuse preferentially towards these sinks where they become stabilized and are then harmless to the device. Such a “gettering” procedure is efficient only under following conditions:

- The metal impurities must be mobile: precipitated impurities or impurity clusters need first to be dissociated before being trapped by gettering sites.
- The impurity diffusion length during the heat treatments which follow the initial contamination step must be larger than the distance between the impurities and the “sinks”.

- The trapping sites must be effective at temperatures where impurity diffusion occurs and avoid impurity release during subsequent heat treatments.

The gettering “efficiency” of a given process, which can be defined as the ratio between the impurity concentrations in the active region of a device with and without gettering sites, depends strongly on the physical properties of the metal impurity (equilibrium configuration, diffusion coefficient, charge state, ...) discussed above. For fast-diffusing impurities such as Cu or Ni, which easily form silicide particles, the first condition may constitute a limiting factor, while for metals such as Fe or Au, which most frequently are dissolved on interstitial or substitutional sites, respectively, no preliminary dissociation step is required. Conversely, the second condition is of significant importance for the slowest-diffusing species.

The heat treatment required for the gettering process to take place depends on the nature of the sinks. For processes described as “segregation-type gettering”, the sink corresponds to a region characterized by a larger metal solubility compared with standard silicon. Highly doped regions or aluminum thin films deposited on the back surface of the silicon are some popular examples [48, 49]. In the former case, p^+ doping, for instance, enhances the solubility of positively charged impurities (such as interstitial Cu or Fe). In epitaxial p/p^+ wafers, the entire p^+ substrate corresponds to the gettering region. This solution is particularly well suited to microelectronic devices. In the second example, the aluminum–silicon eutectic formed at the Al/Si interface has metal solubilities orders of magnitude larger than those of silicon and constitutes the gettering layer. This technology has been used successfully to purify silicon crystals used for photovoltaic applications. In both cases, impurity gettering occurs during the anneal provided that the temperature is high enough to fulfill the second condition described above. The gettering efficiency is strongly dependent on the segregation coefficient between silicon and the gettering layer and on the impurity diffusion coefficient at the annealing temperature.

In another important class of gettering processes, known as “relaxation-type gettering”, the sinks act as nucleation centers for metal silicides and tend to reduce any metal supersaturation [50]. Thus, during cooling, precipitation of supersaturated metal impurities occurs preferentially at the gettering sites. This introduces a concentration gradient, which sustains metal diffusion towards the sink. The gettering efficiency depends on the supersaturation level and on the impurity diffusion length. The supersaturation increases with the cooling rate and triggers precipitation. The effective impurity diffusion length, on the other hand, decreases with increasing cooling rate and must remain large enough to allow the impurities reach the growing silicide particles. High efficiencies are thus possible only in a relatively narrow process window. The most frequently used method, also called intrinsic gettering, uses oxygen clusters and related extended defects (dislocations), formed in Czochralski silicon during a high-temperature process, as nucleation sites for heterogeneous

metal precipitation. By previously outdiffusing oxygen from the near-surface region, one causes oxygen precipitates to form only beyond the active regions of devices and to efficiently getter metals during subsequent anneals [50, 51].

Defects induced by ion implantation constitute another way to introduce efficient gettering sinks in the proximity of the active regions of devices [52–54]. Extended defects, such as dislocations or stacking faults, generally constitute nucleation sites for heterogeneous metal precipitation (similarly to oxide particles). Nanometer-sized cavities formed by high-dose He or H implantation have been found to efficiently trap metal impurities on the cavity walls. In this case, the gettering mechanism can be described in terms of metal segregation towards the cavities with a segregation coefficient which decreases with increasing cavity wall coverage [53].

Gettering of various metal impurities has also been found to occur efficiently during phosphorus diffusion in silicon [55]. Segregation of acceptor-type impurities (Cu_{sub} , for instance) to the highly doped n-type region and formation of P–metal pairs have been suggested as possible driving forces for the gettering mechanism [49]. In this process, injection of point defects (mostly self-interstitials) induced by P diffusion also plays a dominant role, since self-interstitials enhance metal diffusion (kick-out mechanism) and facilitate the dissolution of metal-rich silicide particles. Some reports have found that a combination of phosphorus diffusion with aluminum back-surface gettering is an effective gettering solution, especially for solar cell applications, for which both steps are required for the formation of the device structure [56].

In recent decades, impurity gettering has been an active research topic, with the final aim of establishing a global physical model which could be used by a device manufacturer to forecast and optimize the gettering efficiency of its technology. These models rely on the fundamental properties of metal impurities discussed in the previous sections. The state of the art of gettering modeling has been described in several recent reports [57–59]. Some issues such as the stability of the gettering sites (condition 3) and the kinetics of precipitate dissolution (condition 1) are still under investigation.

13.4.2 Trace Detection

For ultralarge-scale integrated devices the tolerable metal impurity concentrations, of the order of 10^{10} to 10^{12} cm^{-3} , are below the detection limit of most standard analytical tools, such as secondary-ion mass spectroscopy and atomic absorption spectroscopy. Highly sensitive methods such as total X-ray fluorescence spectroscopy and inductively coupled plasma mass spectroscopy either are limited to surface analyses or require sample destruction and are time-consuming. Electrical methods constitute an interesting alternative, with the potential for low detection limits and contactless nondestructive procedures. The reliability of these methods depends strongly on our understanding of the properties discussed above of metal impurities in silicon and on our understanding of their interaction with charge carriers. For instance,

traces of iron in silicon can be detected efficiently by measuring changes in the minority carrier lifetime upon dissociation of iron–boron pairs. This can be achieved on whole wafers in a contactless mode by associating exposure to high-intensity light (to promote pair dissociation through a recombination-enhanced reaction) with standard equipment for the measurement of minority carrier lifetime or diffusion length. The resulting detection limit lies below 10^{10} atoms per cm^3 [44, 60].

Similar attempts to detect traces of copper in silicon by minority carrier diffusion length measurements have been reported recently by Yli-Koski et al. [61]. These authors found that exposure to high-intensity light enhances the precipitation rate of interstitial copper on extended defects (oxide particles) and induces an irreversible drop in the minority carrier lifetime. Another methodology is based on the extremely high mobility of positively charged interstitial copper [62–64]. Electrical capacitive transient signals are delivered by simple devices (Schottky barriers) if the electric field in the depletion region allows interstitial copper atoms to experience drift. These signals can then be used to detect quantitatively the presence of as few as 10^{11} copper atoms per cm^3 of silicon. However, for such a measurement to be significant, most of the copper impurities need first to be put into the interstitial sites by an appropriate heat treatment. Experimental investigations of the physical properties of copper in silicon have shown that this condition can be fulfilled in silicon of high crystalline quality by short anneals at temperatures as low as 600°C [62].

13.4.3 Other Engineering Issues

Other engineering topics related to metal impurities include defect passivation and minority lifetime engineering. The former uses defect reactions between an electrically active defect and a mobile impurity to form complexes whose electrical activity is reduced or “passivated”. The most extensively investigated example of a mobile passivating impurity is given by hydrogen in silicon [65]. Passivation of extended defects in polycrystalline silicon materials has become a widespread technique to enhance the energy conversion efficiency of solar cells [66]. The presence of hydrogen atoms generally reduces the recombination of excess minority carriers at silicon grain boundaries. Hydrogen has also been observed to form pairs with most metal impurities (see Chap. 14 in this book). However, some of these introduce deep levels which differ from those specific to the isolated metal impurity and hence remain electrically active. Since hydrogen is readily introduced into silicon at room temperature during wet chemical processing, hydrogen-related defect reactions should always be considered when characterizing the electrical activity of a given impurity.

Minority lifetime engineering is related to the development of fast switching devices for which the carrier lifetime controls the allowed switching rate.

The introduction of efficient recombination centers helps to lower the corresponding time constant. Gold and platinum impurities introduce deep levels, which make them useful for this task [37]. The final choice of the engineer includes other related aspects such as the generation activity (responsible for leakage currents), impurity diffusion and solubility.

13.5 Conclusion

The numerous experimental and theoretical works on metal impurities in silicon published so far have led to a high level of understanding of defects never achieved for other semiconductor materials. They have certainly contributed to the rapid evolution of the microelectronics industry over recent decades. Yet, with the continuous downscaling of microelectronic device dimensions, the tolerance for metal contamination is still shrinking. The formation of shallow junctions requires lower annealing temperatures, which may be incompatible with standard gettering solutions. Contamination monitoring requires lower detection limits. Also, new materials will be introduced into silicon devices for dielectrics, interconnects or diffusion barriers and introduce a risk of contamination by less familiar impurities. We may thus conclude that the physics of metal impurities in silicon, even after 50 years of active research, will remain an important issue in the near future.

References

1. W. Bergholz, D. Gilles: *Phys. Stat. Sol. (b)* **222**, 5 (2000)
2. E.P. Burte, W. Aderhold: *Sol. Stat. Electron* **41**, 1021 (1997)
3. R. Gonella, P. Motte, J. Torres: *Microelec. Reliab.* **40**, 1305 (2000)
4. K.S. Low, M. Schwerd, H. Koerner, H.J. Barth, A. O'Neil: *SPIE Conf. Vol. 3883* (1999) p. 24
5. W. Schröter, J. Kronewitz, U. Gnauert, F. Riedel, M. Seibt: *Phys. Rev. B* **52**, 13726 (1995)
6. A. Huber, G. Bohm, S. Pahlke: *J. Radioanal. Nucl. Chem.* **169**, 93 (1993)
7. K. Graff: *Metal Impurities in Silicon-Device Fabrication*, 2nd edn (Springer, Berlin, Heidelberg 2000)
8. E.R. Weber: *Appl. Phys. A* **30**, 1 (1983)
9. A.G. Milnes: *Deep Impurities in Semiconductors* (Wiley, New York 1973)
10. M. Lannoo, J. Bourgoin: *Point Defects in Semiconductors II* (Springer, Berlin, Heidelberg 1989)
11. R.H. Hall, J.H. Racette: *J. Appl. Phys.* **35**, 379 (1964)
12. G.W. Ludwig, H.H. Woodbury: *Solid State Phys.* **13**, 223 (1962)
13. H. Lemke: *Phys. Stat. Sol. (a)* **1**, 283 (1970)
14. T. Heiser, A.A. Istratov, C. Flink, E.R. Weber: *Mater. Sci. Eng. B* **58**, 149 (1999)
15. A.A. Istratov, C. Flink, H. Hieslmair, E.R. Weber, T. Heiser: *Phys. Rev. Lett.* **81**, 1243 (1998)

16. J. Utzig: *J. Appl. Phys.* **65**, 3868 (1989)
17. A.A. Istratov, H. Hieslmair, E.R. Weber: *Appl. Phys. A* **70**, 489 (2000)
18. T. Prescha, T. Zundel, J. Weber, H. Prigge, P. Gerlach: *Mater. Sci. Eng. B* **4**, 79 (1989)
19. D.E. Woon, D.S. Marynick, S.K. Estreicher: *Phys. Rev. B* **45**, 13383 (1992)
20. U. Gösele, F. Morehead, W. Frank, A. Seeger: *Appl. Phys. Lett.* **38**, 157 (1981)
21. H. Bracht, H. Overhof: *Phys. Stat. Sol. (a)* **158**, 47 (1996)
22. H. Bracht: *Defects Diffusion Forum* **143–147**, 979 (1997)
23. T. Heiser, A. Mesli: *Phys. Rev. Lett.* **68**, 978 (1992)
24. H.P. Gunnlaugsson, G. Weyer, M. Dietrich, M. Fanciulli, K. Bharuth-Ram, R. Sielemann: *Appl. Phys. Lett.* **80**, 2657 (2002)
25. A.G. Cullis, L.E. Katz: *Phil. Mag.* **30**, 1419 (1974)
26. M. Seibt, K. Graff: *J. Appl. Phys.* **63**, 4444 (1988)
27. M. Seibt, W. Schröter: *Phil. Mag. A* **59**, 337 (1989)
28. T. Heiser, S. McHugo, H. Hieslmair, E.R. Weber: *Appl. Phys. Lett.* **70**, 3576 (1997)
29. S.A. McHugo, A. Mohammed, A.C. Thompson: *J. Appl. Phys.* **91**, 6396 (2002)
30. H. Feichtinger, J. Ostwald, R. Czaputa, P. Vogl, K. Winstel. In: *13th International Conference on Defects in Semiconductors*, Coronado 1984, ed. by L.C. Kimerling, J.M. Parsey (Metallurgical Society of the AIME, Warrendale, PA 1984) p. 855
31. A. Chantre, L.C. Kimerling: *Mater. Sci. Forum* **10–12**, 387 (1986)
32. H. Lemke: *Phys. Stat. Sol. (a)* **75**, K49 (1983)
33. H. Lemke: *Phys. Stat. Sol. (a)* **71**, K215 (1982)
34. W. Shockley, W.T. Read: *Phys. Rev.* **87**, 835 (1952)
35. R.N. Hall: *Phys. Rev.* **87**, 387 (1952)
36. S.D. Ganichev, E. Ziemann, W. Prettl, I.N. Yassievich, A.A. Istratov, E.R. Weber: *Phys. Rev. B* **61**, 10361 (2000)
37. O. Bostrom, B. Pichaud, M. Regula, J.C. Bajard, G. Blondiaux, O.A. Soltanovich, E.B. Yakimov, A. Lhorte, J.B. Quoirin: *Mater. Sci. Eng. B* **71**, 166 (2000)
38. A.A. Istratov, H. Hieslmair, C. Flink, T. Heiser, E.R. Weber: *Appl. Phys. Lett.* **71**, 2349 (1997)
39. S.K. Estreicher: *Phys. Rev. B* **60**, 5375 (1999)
40. J. Weber, H. Bauch, R. Sauer: *Phys. Rev. B* **25**, 7688 (1982)
41. T. Prescha: Ph.D. Thesis, Max Planck Institute, Stuttgart, Germany (1990)
42. S. Knacke, J. Weber, H. Lemke, H. Riemann: *Phys. Rev. B* **65**, 165203 (2002)
43. R. Czaputa: *Appl. Phys. A* **49**, 431 (1989)
44. G. Zoth, W. Bergholz: *J. Appl. Phys.* **67**, 6764 (1990)
45. H. Hedemann, W. Schröter: *Solid State Phenom.* **57–58**, 293 (1997)
46. S.A. McHugo, A.C. Thompson, A. Mohammed, G. Lamble, I. Périchaud, S. Martinuzzi, M. Werner, M. Rinio, W. Koch, H.-U. Hoefs, C. Haessler: *J. Appl. Phys.* **89**, 4282 (2001)
47. B.O. Kolbesen, H. Cerva: *Phys. Stat. Sol. (b)* **222**, 303 (2000)
48. R. Hoelzl, K.J. Range, L. Fabry: *Appl. Phys. A* **75**, 525 (2002)
49. E. Spiecker, M. Seibt, W. Schröter: *Phys. Rev. B* **55**, 9577 (1997)
50. D. Gilles, E.R. Weber, S. Hahn: *Phys. Rev. Lett.* **64**, 196 (1990)
51. A. Bazzali, G. Borionetti, R. Orizio, D. Gambaro, R. Falster: *Mater. Sci. Eng. B* **36**, 85 (1996)

52. F. Roqueta, A. Grob, J.J. Grob, R. Jerisian, J.P. Stoquert, L. Ventura: Nucl. Instrum. Methods Phys. Res. B **151**, 298 (1999)
53. S.M. Myers, D.M. Follstaedt: J. Appl. Phys. **79**, 1337 (1996)
54. R. Kögler, J.R. Kaschny, R.A. Yankov, P. Werner, A.B. Danilin, W. Skorupa: Solid State Phenom. **57–58**, 63 (1997)
55. W. Schröter, R. Kühnapfel: Appl. Phys. Lett. **56**, 2207 (1990)
56. M. Loghmarti, K. Mahfoud, J. Kopp, J.C. Muller, D. Sayah: Phys. Stat. Sol. (a) **151**, 379 (1995)
57. J.S. Kang, D.K. Schroder: J. Appl. Phys. **65**, 2974 (1989)
58. S.A. McHugo, H. Hieslmair, E.R. Weber: Appl. Phys. A **64**, 127 (1997)
59. S.M. Myers, M. Seibt, W. Schröter: J. Appl. Phys. **88**, 3795 (2000)
60. L. Kronik, Y. Shapira: Surf. Sci. Rep. **37**, 1 (1999)
61. M. Yli-Koski, M. Palokangas, A. Haarahiltunen, H. Väinölä, J. Storgårds, H. Holmberg, J. Sinkkonen: J. Phys. Condens. Matter **14**, 13119 (2002)
62. A. Belayachi, T. Heiser, J.P. Schunck, S. Bourdais, P. Bloechl, A. Huber, A. Kempf: Mater. Sci. Eng. B **102**, 218 (2003)
63. T. Heiser, S. McHugo, H. Hieslmair, E.R. Weber: App. Phys. Lett. **70**, 3576 (1997)
64. T. Heiser, E.R. Weber: Phys. Rev. B **58**, 3893 (1998)
65. S.K. Streicher, J.L. Hastings, P.A. Fedders: Mat. Sci. Eng. B **58**, 31 (1999)
66. H.E.A. Elgamel, J. Nijs, R. Mertens, M.G. Mauk, A.M. Barnett: Sol. Energ. Mater. Sol. C **53**, 277 (1998)

14 Hydrogen

C.A.J. Ammerlaan

14.1 Introduction

Hydrogen, the most abundant substance in the universe, is also omnipresent on earth in the form of water, the most common liquid, following its name “hydro-gen”. Being the first element in the periodic table of elements, hydrogen can show extreme properties. It is the smallest and lightest element of all. Consequently, hydrogen is frequently present as an impurity in crystals of every kind, and so too for silicon. When it is singly ionised only its nucleus, the bare proton, with dimensions of femtometres, is left. Without any core electrons, hydrogen is a unique impurity. The positively charged impurity will seek a position with maximum electron density, or lowest Coulomb potential, in the crystal, creating bonding in a typically ionic manner. In the neutral charge state, the hydrogen atom will have its 1s electron shell singly occupied. Using a different option, hydrogen can accept a second electron in this low-energy state offered by the 1s shell, which will become completely filled. This full shell is the typical form realised in covalent bonding. With these contrasting bonding schemes, the bonding of hydrogen in silicon is anticipated to be complex in its appearance. One cannot expect valid bonding models to be based on simple intuitive thinking. Rather, this is the domain of the most advanced theoretical computations. Such calculations, possible nowadays and undoubtedly possible with even greater precision in the future, have already shown the vital role of theory in the understanding of the behaviour of hydrogen in silicon. Maybe theory is having a larger impact than ever before on the physics of defects and impurities in semiconductors, with the obvious exception of shallow substitutional donors and acceptors and the celebrated effective-mass theory. Owing to its electronic structure, hydrogen is a highly reactive impurity. These expectations have been fully confirmed by the observation of interactions of hydrogen with a rich variety of impurities, including shallow and deep states, substitutional and interstitial impurities, and donor, acceptor and amphoteric impurities, from all rows and columns of the periodic table. A very special case is the interaction of hydrogen with itself in forming the H_2 molecule. This prototype of covalent bonding creates a molecule without any charge or dipole moment. In this reaction hydrogen has fully passivated its own electrical activity, a feature that hydrogen can also show in interactions with other impurities, such as acceptors and dan-

gling bonds. Several powerful experimental techniques, such as EPR, electron paramagnetic resonance, are not capable of observing passivated products, which creates a handicap for experimental investigations. Though possibly present in great quantities, hydrogen-passivated centres do not have any appreciable effect on material properties and are hard to detect. However, upon heat treatment, e.g. in device processing, passivated complexes may dissociate and reactivate the component impurities. In this way hydrogen can be responsible for unstable properties of silicon.

In the past two decades, intensive research on the effects of hydrogen in silicon has been conducted. Several methods for controlled intentional introduction of hydrogen have been developed, each with its own advantages and limitations. By exposing silicon wafers downstream from a plasma source of hydrogen atoms, silicon can be hydrogen-doped to a large concentration in a thin surface layer, typically 1 µm thick. Substrates are usually kept at an elevated temperature around 500°C. Another way to obtain shallow highly doped layers is chemical etching at room temperature, or even boiling in water, of silicon wafers. Treating samples in a wet atmosphere is a typical condition that also arises in standard IC processing. As a result, hydrogen can be introduced unintentionally and, if this is not realised, it can be responsible for unexpected phenomena and properties. A controlled introduction can be achieved by proton implantation. The range of introduction will depend on the implantation energy and can be up to several centimetres. The concentrations achieved are determined by the implantation dose and can vary widely as well. Implantation is accompanied by creation of radiation damage, with the usual requirement for thermal annealing of these defects. Diffusion at high temperatures, near 1200°C, for several hours can give uniform doping over millimetre distances.

For experimental research purposes, the method of hydrogen introduction selected will depend on the method of measurement envisaged. Shallow layers created by etching provide ideal samples for measurements by deep-level transient spectroscopy (DLTS). This sensitive technique yields concentrations and energy levels. Structural information is better obtained by spectroscopy in the infrared of vibrational modes of complexes involving the light hydrogen impurity. With two stable isotopes, the proton and deuteron, with a relative mass difference of a factor two, larger than for any other element, the identification of hydrogen modes can be safely performed. Local vibrational mode (LVM) spectroscopy is possible for all charge states of the defects. Even though this is not the case for magnetic resonance, this latter technique is still very suitable for detection and characterisation of hydrogen-related complexes if they are not fully passivated. Again, the availability of two isotopes is of great advantage for the identification of hydrogen in the centres. The different nuclear spins, $I_p = 1/2$ for the proton and $I_d = 1$ for the deuteron, give a clear distinction in the electron paramagnetic resonance (EPR) spectra. When electron–nuclear double resonance (ENDOR) can be performed, the different

nuclear magnetic moments, $\mu_p = +2.79285 \mu_N$ and $\mu_d = +0.85744 \mu_N$, allow an unambiguous identification of hydrogen. The crucial complementary role of theory has already been stressed.

Several reviews of hydrogen in silicon have been published in recent years [1–14]. These reviews include references to over a thousand original papers in the field, allowing one to follow the investigations in all details. This brief review will be limited to the discussion of four specific cases. First, single hydrogen atoms and molecules in silicon, as the basic structures, will be considered. The passivation of acceptor and donor impurities, is taken as the second and third examples. Finally, the interaction of hydrogen with transition metal impurities, as typical deep-level centres, is described.

14.2 Hydrogen Atoms and Molecules

Hydrogen is an amphoteric impurity in silicon. Depending on the Fermi level, as determined by the shallow dopants, it will assume either a positive or a negative charge state, or it may be neutral under nonequilibrium conditions. As an ionised donor, in the state H^+ , hydrogen will find its energetically most stable position in the region of highest electron density. This will be midway between two neighbouring silicon atoms, where the covalent bond between these atoms creates the highest electron density. Also, the neutral hydrogen atom will be bonded at this so-called bond-centred site, referred to as the BC site [15–23]. The negative hydrogen ion H^- [24–26], on the other hand, will prefer the region of lowest electron density, which is found in an interstitial space at the T site [20, 22]. The established structure models are given in Fig. 14.1 [27]. With hydrogen in the BC position, the two silicon neighbour atoms in the Si–H–Si three-centre bond move outwards from the hydrogen, possibly in an asymmetric way, to minimise the energy [15, 17, 19, 21, 22, 28–30]. Theoretically calculated distances are indicated in Fig. 14.1 [22, 27]. With the positive hydrogen ion situated in a region of high electron density, it is difficult to add one more electron to convert the impurity to the neutral state. In other words, a high Fermi level is required to create the neutral state, and the (+/0) donor level will be high in the bandgap. Using capacitance–voltage techniques, such as deep-level transient spectroscopy, the donor level of single hydrogen was determined as $E_d = E_c - 0.2 \text{ eV}$ [31–33]. In the earlier literature, this level, not yet identified as hydrogen-related, was labelled the E3' centre [34]. For negative hydrogen, firmly bonded on the T site with low electron density, similar arguments, but with opposite parameters, predict a lower energy for the level (0/–), where the centre will lose its electron. Experiment has determined this level to be $E_a = E_c - 0.56 \text{ eV}$, near the midgap position [35–37]. Both ions tend to be very stable, owing to the attractive electrostatic potentials in which they are accommodated. As such an option does not exist for hydrogen in its neutral state, the energy for H^0 is expected to be higher. In Fig. 14.2 [22], the energies for creation of H^- and H^+ ions

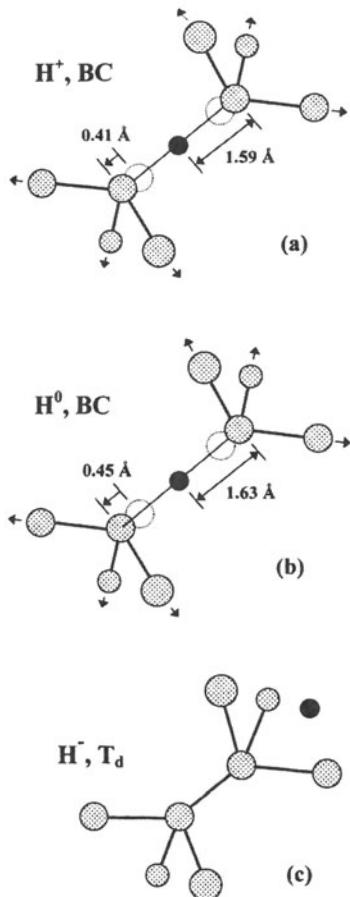


Fig. 14.1. Schematic illustrations of the location of (a) H^+ , (b) H^0 and (c) H^- in the silicon crystal. Relaxations of the silicon atoms, based on calculations presented in [22], are indicated. After Johnson and Van de Walle [27]

relative to the neutral H^0 are presented as a function of the Fermi energy, the energy which has to be transferred in an exchange of an electron with a reservoir [20–22, 29, 38]. As argued, the acceptor level $E_a(0/-)$ is below the donor level $E_d(0/+)$, which is reminiscent of centres with a negative correlation energy U . In the present case $U = E_a - E_d = -0.36 \text{ eV}$ [35], but the large energy gain is due to the different lattice sites occupied by hydrogen in its oppositely ionised states. This feature distinguishes the case of hydrogen from documented negative- U systems in silicon, such as the lattice vacancy and the interstitial boron impurity. In these latter cases energy is gained from Jahn–Teller-driven lattice distortions. Inspection shows that the energy

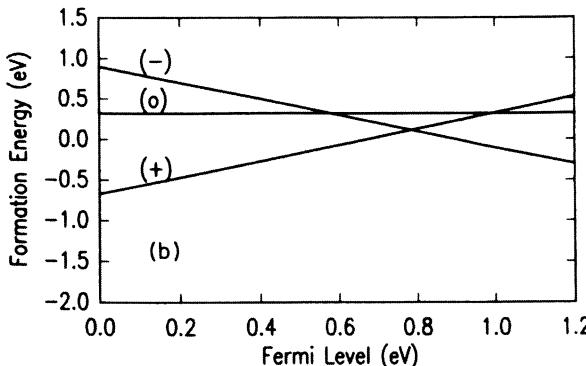


Fig. 14.2. Qualitative indication of the relative stability of different charge states of a hydrogen interstitial impurity in silicon. After Van de Walle et al. [22]

of H^0 is never the lowest, with the conclusion that H^0 is an unstable species. Neutral hydrogen atoms can lower their energy by following the option $2H^0 \rightarrow H^+ + H^-$. Transitions between lattice sites forced by changes of charge state in the space charge region of a p–n junction have been observed [32]. The barrier height between BC and T sites was measured to be 0.29 eV. The transition rates for the proton and deuteron follow the square root of the mass ratio, indicating that the process just involves the jump of one hydrogen atom.

In its neutral charge state the hydrogen atom, with its one unpaired electron, is paramagnetic and has been observed in magnetic resonance [30, 33, 39–41]. As H^0 is unstable, illumination of the sample is required to produce the centre and its EPR spectrum, labelled Si-AA9. An important feature of the spectrum, as shown in Fig. 14.3 [40], is the doublet structure of the main line, which is the result of hyperfine interaction with one proton, nuclear spin $I_p = 1/2$. In deuterated material, the splitting is threefold since $I_d = 1$, with a smaller splitting corresponding to the smaller nuclear magnetic moment of the deuteron. From the angular dependence, the centre symmetry has been determined as trigonal, providing confirmation of the BC site model. A close correspondence has been established between these EPR experiments on hydrogen and muon spin rotation studies of the positive anomalous muon and of the muonium atom [42–46]. The muon can be considered as a light isotope of hydrogen, with identical bonding behaviour in crystals of the diamond structure [47]. Table 14.1 summarises data for the two systems obtained by experiment and theory [41, 43, 48–51]. It has been concluded that hyperfine interactions with the proton and muon scale in proportion to the nuclear moments, giving a factor 3.17 for μ_μ/μ_p . The hyperfine interactions with the two equivalent silicon nearest-neighbour atoms in the Si– μ /p–Si bond are practically equal. This demonstrates the equivalence of the two cases and renders the muon a valuable substitute for hydrogen studies without the handicap of

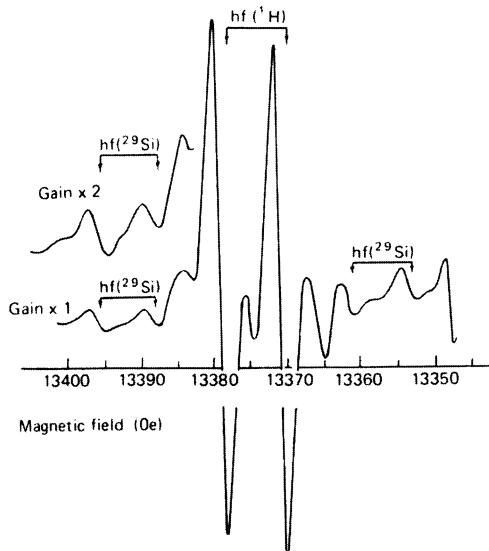


Fig. 14.3. Electron paramagnetic resonance spectrum Si-AA9 of neutral hydrogen in silicon. Hyperfine interactions with ^1H and ^{29}Si are indicated by arrows. Observation conditions: sample illuminated, temperature 77 K, magnetic field parallel to $\langle 100 \rangle$, microwave frequency 37.47 GHz. After Gorelkinskii and Nevinnyi [40]

Table 14.1. Hyperfine parameters, obtained from EPR and μ SR experiments and from spin-density functional theory, for the centre $(\text{H}_{\text{BC}})^0$ and anomalous muonium, Mu^* . After Ammerlaan and Huy [52]

Centre	Nucleus	$A_{ }$ (MHz)	A_{\perp} (MHz)	a (MHz)	b (MHz)	Method	Reference
$(\text{H}_{\text{BC}})^0$	^1H	-6.2	-31.4	-23.0	8.4	EPR	[41]
Mu^*	Mu	-16.82	-92.59	-67.33	25.26	μ SR	[50]
Mu^*	Mu	9.6	-57.3	-35	22.3	Theory	[51]
$(\text{H}_{\text{BC}})^0$	^{29}Si	-139.0	-72.9	-94.9	-22.0	EPR	[41]
Mu^*	^{29}Si	-137.5	-73.96	-95.1	-21.2	μ SR	[43]
Mu^*	^{29}Si	-128	-63.5	-85	-21.5	Theory	[51]

requiring a paramagnetic state. The anisotropic part of the hyperfine interaction with the proton is understood as a dipole-dipole interaction with the electron spin density localised on the neighbouring silicon atoms. To account for the observed magnitude $b = 8.4 \text{ MHz}$, a distance between the proton and the silicon of 1.65 \AA is required [52]. Compared with the regular parameters of silicon, an outward relaxation by 0.47 \AA is thus derived, in excellent agreement with theoretical results for this relaxation [27]. Both neutral and positive hydrogen are trigonal centres and are oriented along one of the four available

$\langle 111 \rangle$ crystal axes, with equal probability in the random case. Applying a uniaxial stress, in a suitably chosen direction, at a temperature where the hydrogen can jump between adjacent BC sites will induce a preferential alignment [30]. In the EPR spectrum, the concentrations of centres in the various distinct orientations can be monitored by the amplitudes of the corresponding resonances. Analysis of the repopulation in terms of a piezospectroscopic tensor for the trigonal case gives a negative energy per unit strain, confirming the outward relaxation of the silicon atoms with respect to the hydrogen [30]. From isochronal annealing studies carried out with samples in darkness, the reorientation process for H^+ was found to follow first-order kinetics and an Arrhenius temperature dependence in the range 120 to 150 K. The frequency for jumps between BC sites is given by $\nu(T) = \nu(\infty) \exp(-E_a/kT)$, where $\nu(\infty) = 2.3 \times 10^{12} \text{ s}^{-1}$ and $E_a = 0.43 \text{ eV}$. It has thus been found that the energy barrier for reorientation, involving an elementary jump from one BC site to a neighbouring one, is essentially equal to the energy $E = 0.48 \text{ eV}$ observed for the permeation by positive hydrogen of a 2 mm thick wall at elevated temperatures in the range 1092–1200°C [53]. Figure 14.4 [54] summarises results obtained for positive-hydrogen diffusion by experimental and theoretical methods [29, 53, 55–57]. It can be concluded that a consistent

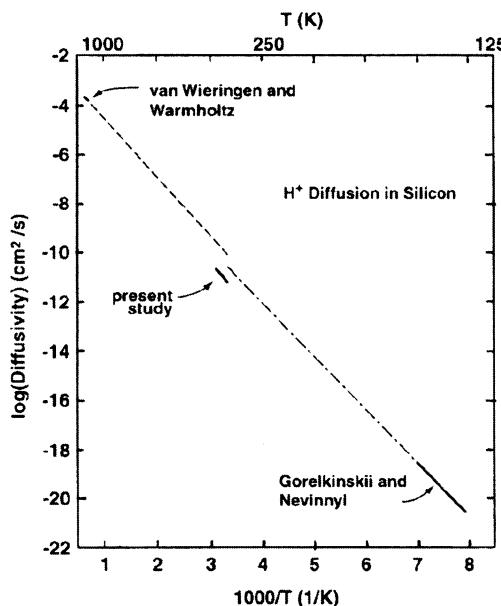


Fig. 14.4. Diffusivity of hydrogen H^+ in silicon from measurements at high temperatures by van Wieringen and Warmoltz [53], at low temperatures by Gorelkinskii and Nevinnyi [30] and at room temperature by Herring et al. [54]. After Herring et al. [54]

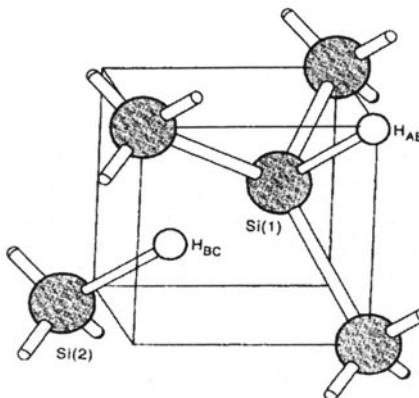


Fig. 14.5. Atomic structure of the H_2^* defect. After Holbech et al. [59]

set of data exists over an impressive range of temperatures and diffusion rates [54]. The solubility of atomic hydrogen at high temperatures is also given by an Arrhenius-type expression $S(T) = S(\infty) \exp(-E_s/kT)$, where $S(\infty) = 4.8 \times 10^{21} \text{ cm}^{-3}$ and $E_s = 1.87 \text{ eV}$ [53].

As in vacuum, H_2 molecules represent a stable form of hydrogen inside silicon as well. The molecule straddles the T site with almost degenerate energies for orientations along $\langle 100 \rangle$ and $\langle 111 \rangle$ but experiences a high barrier for diffusion [17, 20, 22, 23, 29, 58]. Another geometrical configuration of the hydrogen dimer consists of hydrogen atom near a BC site and a second atom on an AB site. This H_2^* centre is illustrated in Fig. 14.5 [59]. It can be considered as being formed from the attraction of a single H^+ ion on the BC site and an H^- ion on the T site [21, 23, 60–62]. It has a slightly higher formation energy than the molecule but diffuses more rapidly. As these forms of hydrogen have no gap levels and no dipole moment they are difficult to detect experimentally, and much of the modelling has come from theory. In experiments, local vibrational mode spectroscopy has been the most informative technique, both for the monatomic species and for the dimer, and has contributed to confirmation of the structural models. For H^+ on the BC site, the reported stretch frequencies of the Si–H bond are 1990 cm^{-1} for the proton and 1440 cm^{-1} for the deuteron, scaling properly with mass [22, 63–65]. For the dimer H_2^* , the different frequencies observed for the Si–H stretch modes confirm the presence of a pair of inequivalent hydrogen positions in the defect [59, 62, 66]. Shifts induced by H/D isotope substitution can be interpreted consistently in the scheme of defects $\text{H}_{\text{BC}}-\text{H}_{\text{AB}}$, $\text{H}_{\text{BC}}-\text{D}_{\text{AB}}$, $\text{D}_{\text{BC}}-\text{H}_{\text{AB}}$ and $\text{D}_{\text{BC}}-\text{D}_{\text{AB}}$ and follow quite closely the square-root-of-mass dependence. Line splittings induced by uniaxial stress confirm the trigonal symmetry of the centre. In addition, the experimental frequencies are in good agreement with ab-initio local-density-functional cluster calculations.

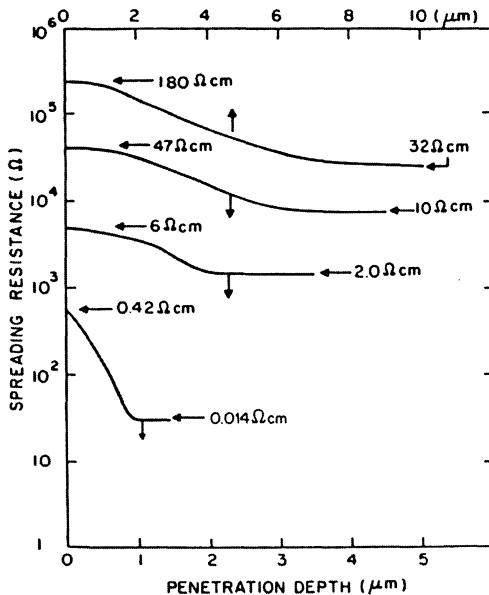


Fig. 14.6. Spreading-resistance profiles of four boron-doped silicon samples hydrogenated at 122°C for 1 h for the lowest resistivity and 4 h for the other three resistivities. Note the different lower and upper depth scales and the greater penetration depths of atomic hydrogen for the lower boron concentrations. After Pankove et al. [69]

14.3 Passivation of Acceptors

The passivation, or neutralisation, of acceptor states has drawn attention strongly to the field of hydrogenation and to its drastic effects on semiconductor materials and device properties [67, 68]. Passivation of single substitutional acceptors in bulk single-crystalline silicon by ionised hydrogen donors H^+ is observed in its most straightforward form as an increase of the resistivity in p-type material, as illustrated in Fig. 14.6 for the acceptor boron [69]. Passivation manifests itself also by a decrease of the free-carrier concentration [70, 71], a decrease of the photoluminescence of acceptor-bound excitons [72], the appearance of new local vibrational modes [70, 73–78] and an increase of carrier mobility [70, 71]. This latter effect distinguishes passivation from compensation. In the case of compensation, spatially unrelated donors and acceptors are present as dopants of opposite type, leading to a hole concentration $n(A^-) - n(H^+)$ and a concentration of ionised impurity scattering centres $n(A^-) + n(H^+)$. With passivation, the concentration of free carriers will be equal but the concentration of ionised impurities will be $n(A^-) - n(H^+)$. At temperatures where ionised-impurity scattering governs the mobility (generally low temperatures), the passivation will thus enhance

the mobility. Passivation follows compensation through the attraction between negative substitutional acceptor impurities A^- and positive mobile hydrogen donors H^+ , but leads to new neutral centres which can be viewed, in the first instance, as donor-acceptor pairs. In this concept the donor level of hydrogen is raised into the conduction band by the nearby presence of the negatively charged acceptor. Likewise, the ionisation level of the acceptor is pushed to lower energy into the valence band by the repulsive force of the positive hydrogen donor. The bandgap is swept clean of levels, which is the essential feature of passivation. Most research has been performed on the shallow acceptor boron, with ionisation energy 45 meV, but neutralisation occurs likewise for the other group III acceptors Al, Ga, In and Tl, which have increasingly deeper acceptor levels [69, 72].

With hydrogen migrating in p-type silicon as H^+ via bond-centred sites BC, the most likely structure for a BH complex is an H^+ ion trapped on a BC site next to an acceptor A^- . This structure of a hydrogen inserted between the acceptor and a nearest-neighbour silicon atom along a bonding $\langle 111 \rangle$ direction has received much support from extensive theoretical and experimental investigations [21, 73–75, 78–99]. The defect model is given in Fig. 14.7a [84], with Fig. 14.7b detailing the relaxations taking place to accommodate the extra atom [99]. Numerical values of the bond lengths and displacements obtained from various theoretical calculations and from experiment are given in Table 14.2 [81, 82, 85, 86, 90, 93, 99–102]. One may note that in some results the Si-H distance is shorter than the B-H distance, indicating a structure where the hydrogen passivates the dangling bond on the silicon atom and trivalent boron forms more planar bonds with three silicon neighbours. Both parts are

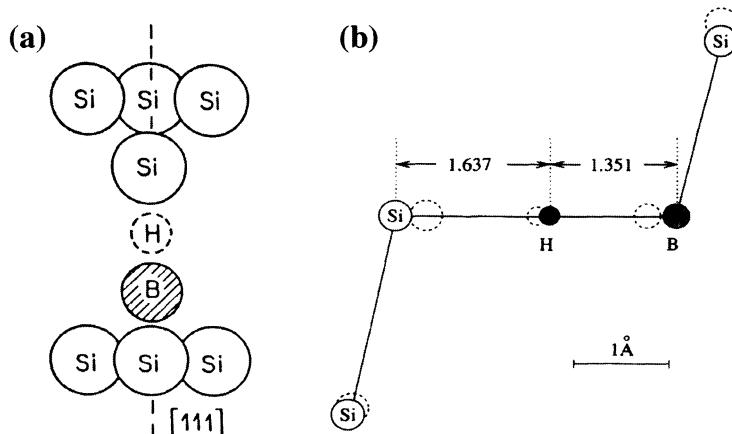


Fig. 14.7. (a) Schematic model of the B-H complex in silicon. After Bergman et al. [84]. (b) Equilibrium configuration of the B-H centre. *Dashed circles* indicate unrelaxed positions. After Zhou et al. [99]

Table 14.2. Equilibrium geometries of the Si–H–B complex in silicon from various calculations. Si_s and B_s indicate positions in an undistorted silicon lattice. All distances in units of Å. In experiments, the boron displacement $\text{B}–\text{B}_s$ was measured to be 0.22 Å [82], 0.28 Å [86], or 0.3 Å [102]. After Zhou et al. [99]

H–Si	H–B	$\text{Si}–\text{Si}_s$	$\text{B}–\text{B}_s$	Reference
1.56	1.51	0.16	0.55	[100]
1.46	1.59	0.22	0.48	[81]
1.44	1.66	0.26	0.49	[93]
1.49	1.53	0.16	0.47	[101]
1.63			0.47	[85]
1.65	1.36	0.24	0.42	[90]
1.64	1.35	0.33	0.31	[99]

fully covalently saturated and have no electrical activity. In other calculations the hydrogen atom is moved more towards the boron atom and a three-centre bond is formed. There exists general, but not unanimous, agreement that the hydrogen in the BH centre is on the $\langle 111 \rangle$ axis, forming a centre of trigonal symmetry, even though the energy surface in directions perpendicular to $\langle 111 \rangle$ is very flat [84, 93, 102–104]. For the acceptors Al, Ga and In, with larger covalent radii, the hydrogen atom is pushed off the axis [84, 93, 95, 105–108]. In an alternative model, the hydrogen atom is placed on the antibonding site of the acceptor, on the $\langle 111 \rangle$ axis, near the interstitial T site, the position of which is indicated in Fig. 14.1c [106, 109–111]. However, in most other calculations this site is found to have a larger energy than the BC site and to form a saddle point on the total-energy surface [80, 88, 90, 91, 93]. Experimental results obtained by perturbed γ – γ angular correlation suggest that hydrogen can be found in both positions [112–114]. Also, the back-bonding site of a silicon neighbour of the acceptor has been suggested as a suitable hydrogen position, similar to what is shown in Fig. 14.8a for the AsH complex [115]. Most of the quantitative modelling is based on theoretical study; experiment has yielded information on the site geometry by alpha particle channelling and nuclear-reaction studies [82, 83, 86, 114]. The experimentally most fruitful technique for studying these passivated centres has been optical absorption related to local vibrational modes [70, 74, 77, 116]. Uniaxial stress has been applied to induce level splitting and draw conclusions about the symmetry, and to induce defect reorientations and study defect kinetics [84, 117]. Dichroism has been used to uniquely monitor specific orientations. For the reorientation of the BH centre, when the hydrogen atom jumps from one BC site to another neighbouring site of the same boron atom, a potential barrier of 0.19 eV was measured [87, 117], in excellent agreement with the activation energy of 0.22 eV obtained in an anelastic relaxation experiment [118] and with calculated potential profiles [88, 90, 91]. A deviation from Arrhenius behaviour observed at low temperatures was related to the process of thermally assisted tunnelling [119, 120]. Isotope effects have been observed that

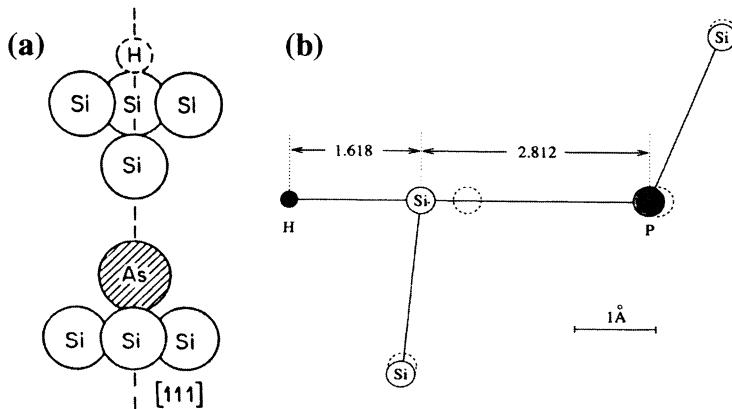


Fig. 14.8. (a) Schematic model of the AsH complex in silicon as an example of the structure of donor-hydrogen complexes. After Bergman et al. [84]. (b) Equilibrium configuration of the P-H centre. *Dashed circles* indicate unrelaxed positions. After Zhou et al. [99]

unequivocally verify the model of one acceptor and one hydrogen atom. The high-resolution results for the stretch mode vibrations are 1905.2 cm^{-1} for $^{10}\text{B}-^1\text{H}$, 1904.4 cm^{-1} for $^{11}\text{B}-^1\text{H}$, 1393.9 cm^{-1} for $^{10}\text{B}-^2\text{H}$ and 1390.6 cm^{-1} for $^{11}\text{B}-^2\text{H}$, and provide severe test data for the theory [78,111]. The remarkably strong boron isotope effect in the deuterated complex was explained as the result of a Fermi resonance [96]. The local-mode phonons due to the stretch mode and the less-studied lower-energy wagging modes [108,121] were also observed in the Raman spectra [76,102–104].

The acceptor-hydrogen complexes are quite stable centres, with binding energies of 1.28 eV for BH , 1.44 eV for AlH , 1.40 eV for GaH and 1.42 eV for InH between the two elements [95]. With respect to the H_2 molecule, the energy per hydrogen atom is 0.6 eV [21,85] or 0.3 eV [23]. The thermal annealing of complexes depends sensitively on the external conditions, such as the illumination level, the carrier concentration and the presence of hydrogen traps, but typically occurs near 150°C [72,76,112,113,122–125].

Double acceptors offer the interesting option of partial or full passivation, probably correlated with the binding of one or two hydrogen atoms to the dopant impurity. Experiments on beryllium have demonstrated tunnelling motion of one hydrogen around the acceptor atom, leading to a complex ground state and complex optical spectra. From experiment it was concluded that the energy minima for hydrogen are in the $\langle 111 \rangle$ directions, as for single acceptors, leading to a tetrahedral rotor [97,126–128]. Theory has indicated energy minima in $\langle 100 \rangle$ directions and consequently an octahedral rotor [89,91]. Similar data for the double acceptor cadmium, provided by perturbed-angular-correlation spectroscopy, were interpreted as thermally

activated tunnelling with an activation energy of 0.21 eV [129]. For the double acceptor zinc, the experiments have identified optical spectra for Zn itself, and for the hydrogen complexes ZnH and ZnH₂ [130–133]. In deep-level transient spectroscopy, the effective passivation of zinc acceptors, reducing their concentration by at least two orders of magnitude, has been observed [134]. It was shown that thermal annealing at 470°C regenerates the acceptors back to their original concentration, with a reactivation energy of 2.2 eV.

14.4 Passivation of Donors

Confirming its power as a universal passivation agent, hydrogen is also capable of neutralising dopants of donor character. It has to be noted, however, that the passivation of the shallow donors phosphorus, arsenic and antimony in silicon is less effective, compared with the figures quoted for acceptors, with maximum donor passivation efficiencies reported as 80% [135, 136]. Donor activity can be reactivated by thermal annealing around 150°C, with binding energies of 1.32 eV for PH, 1.43 eV for AsH and 1.43 eV for SbH [135, 136]. The energy of a hydrogen impurity bound to a donor is only 0.1 eV lower than in the hydrogen molecule H₂, explaining the low stability [85]. Again applying a simple donor–acceptor pairing model, the negatively charged hydrogen ion H[−], occupying interstitial T sites, is attracted by the positive donor D⁺. The passivated donor then consists of a donor on a substitutional site and hydrogen on a nearby interstitial site. Such a model is fully confirmed by advanced theoretical analyses, which all agree on the interstitial site being antibonding to a silicon nearest neighbour of the donor [85, 99, 137–141]. The atomic structure of the trigonal centres, with D–Si–H ordering along a ⟨111⟩ axis, is illustrated in Fig. 14.8 [84, 99]. Quantitative information on the relaxation around the centre, given in Table 14.3 [99, 138–140], shows that both the donor and the silicon neighbour move in the same direction, in contrast to some of the earliest results, in which opposite directions of relaxation were calculated [93, 142, 143]. The first experimental evidence of donor passivation was based on an increase of the resistivity and Hall mobility upon hydrogenation of phosphorus- and arsenic-doped silicon [137, 144]. Substantial support for the defect model was provided by infrared absorption spec-

Table 14.3. Equilibrium geometries of the H–Si–P complex in silicon from various calculations. Si_s and P_s indicate positions in an undistorted silicon lattice. All distances in units of Å. After Zhou et al. [99]

H–Si	Si–P	Si–Si _s	P–P _s	Reference
1.69	2.84	0.66	0.18	[140]
1.66		0.59	0.14	[139]
1.65	2.72	0.52	0.14	[138]
1.62	2.81	0.56	0.10	[99]

troscopy through the observation of local vibrational modes [84, 135]. The line splittings under uniaxial stress of the AsH pair confirm the trigonal symmetry [84, 117]. The silicon–hydrogen stretch-mode wavenumbers of 1555 cm^{-1} for PH, 1561 cm^{-1} for AsH and 1562 cm^{-1} for SbH depend hardly at all on the donor mass, supporting a model in which hydrogen is not a direct neighbour of the donor itself. A hydrogen/deuterium isotope effect with a ratio of the vibration frequencies of about 1.37 for both stretching and wagging modes corresponds to an ordinary global minimum in the potential-energy surface. Mössbauer experiments have revealed that Sb donors can bind two hydrogen atoms [136, 145, 146]. Also, on the basis of theoretical calculations, multiple trapping of hydrogen at shallow acceptors and donors, notably by the formation of BH_2 , PH_2 and SbH_2 complexes, has been predicted [141, 147, 148].

The interaction of hydrogen with the chalcogen deep double donors sulphur, selenium and tellurium, leads to partial passivation of those centres, with the formation of shallow single donors. In infrared absorption spectroscopy, typical effective-mass-like excitation spectra were observed and three sulphur–hydrogen donors were identified, with electron binding energies of 92, 135.07 and 135.45 meV [149, 150]. The spectral shifts upon deuterium substitution prove the presence of one hydrogen atom. Taking advantage of the absence of full passivation, also magnetic resonance could also be applied to the new chalcogen–hydrogen complexes in their neutral paramagnetic states. For both sulphur and selenium, two such complexes were described as spin $S = 1/2$ centres, with trigonal symmetry, consisting of one substitutional chalcogen and one or two interstitial hydrogen atoms [151–156]. One of the selenium centres is interpreted in the most straightforward manner as a selenium–dihydrogen complex. By applying electron–nuclear double resonance (ENDOR), the nuclear spins and magnetic moments of the atoms in the centres were determined for the constituent nuclei ^1H or ^2H and ^{33}S or ^{77}Se , resulting in their unambiguous identification. From the hyperfine interactions, the spin-density distributions in the complexes were established as being spatially very extended, consistent with their shallow-donor nature. Models allocating the hydrogen atoms to the possible BC, Si-AB or S/Se-AB sites on the $\langle 111 \rangle$ trigonal defect axis were proposed. The complete passivation of chalcogen donors by one hydrogen atom, with removal of all bandgap levels by better than a factor of 100, was concluded from DLTS [157–159]. The relevant spectra, as shown in Fig. 14.9 [157], demonstrate the absence of new levels, but shallow states may have escaped detection by this technique. Tellurium centres TeH and a tellurium–multi-hydrogen complex were observed by Mössbauer spectroscopy [146]. In the DLTS and EPR experiments, the chalcogen–hydrogen complexes were found to be stable up to 500°C [153, 157, 159–161]. The reactivation energies were determined as 1.61 eV for SH and 1.39 eV for both SeH and TeH [159]. Theoretical ab-initio modelling studies indicate configurations where hydrogen occupies sites antibonding to a nearest-neighbour silicon atom as the most stable

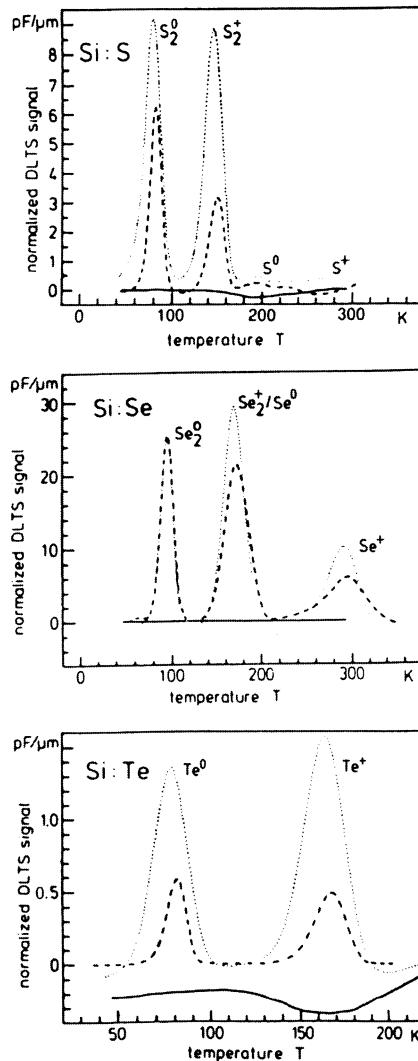


Fig. 14.9. Passivation of the chalcogen elements S, Se and Te, measured by deep-level transient spectroscopy, also showing reactivation of the dopants after thermal annealing at 538°C for 10 min. *Dashed-line spectra* were recorded on a reference sample, *solid curves* after hydrogenation and *dotted curves* after thermal reactivation. After Pensl et al. [157]

complex [141, 162–164]. For sulphur, however, the complex with hydrogen on the BC site has an energy only 0.1 eV higher, and is, taking into account the error limits in the calculations, an additional option. For Se and Te, where the BC site has an energy higher by 0.3 and 0.8 eV, respectively, this form of complex can be excluded. Also, sites antibonding to the chalcogen atom are energetically unrealistic. All the singly hydrogenated complexes have trigonal symmetry and are predicted to be shallow donors. Doubly hydrogenated chalcogen impurities are predicted to be electrically inert. In contrast to an earlier theoretical treatment [165], modern computations provide good agreement with experimental results, in particular those from EPR.

The family of oxygen-related thermal double donors (TDDs) has an unparalleled long history of research aimed at the unravelling of its structure. At an early stage it was concluded from infrared absorption, resistivity measurements and DLTS that these complex donors with an extended core structure are also passivated by hydrogen [166, 167]. In particular, the smaller, earlier species are passivated, resulting in an effective shift of the overall ionisation energy to a shallower level. A nearly complete passivation could be achieved [168]. Thermal annealing in the range 100 to 200°C reactivates the donors, again with the earliest two species showing a behaviour that deviates from that of the later ones [168–171]. Observations made by EPR reveal quite different aspects of hydrogenation. In the EPR spectrum of the centre Si-NL10(H), the presence of hydrogen as a structural component was conclusively demonstrated by ENDOR [172, 173]. A typical spectrum, in which ENDOR transitions were recorded at frequencies symmetrically around the nuclear Zeeman frequency of hydrogen, is shown in Fig. 14.10 [172]. Taking into account the similarities between the Si-NL8 centre, commonly identified with the thermal double donor as defined by the IR absorption spectrum, and the Si-NL10(H) centre, the latter one is most readily interpreted as a hydrogenated thermal donor [174]. But it must be noted that the Si-NL10(H) centre is formed by heat treatment at 450°C, a temperature at which the TDDs passivated at low temperature have already been reactivated [171]. A solution can be found by the acceptance of two different varieties of passivated TDD with hydrogen incorporated at different sites in the complex. With the precise structure of the thermal donor still unknown, the structure of the passivation product is also not yet resolved. It will be a complex centre with low, probably triclinic, symmetry [174]. On the basis of a theoretical study, a partial passivation of the TDDs by binding one hydrogen atom has been reported [175]. An infrared absorption spectrum of one particular family of shallow thermal donors, i.e. STD(H), can possibly be associated with this single donor [176, 177]. Magnetic resonance and infrared absorption have established a link between the presence of STD(H) and Si-NL10(H) centres, suggesting their identity [178].

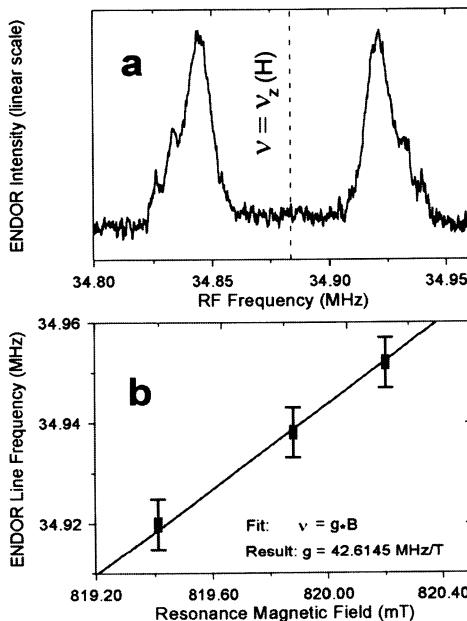


Fig. 14.10. (a) Hydrogen ENDOR spectrum observed after heat treatment at 470°C for 55 h of Czochralski-grown aluminium-doped silicon; (b) shift of the ENDOR frequency with magnetic field, identifying the proton nuclear g factor. After Martynov et al. [172]

14.5 Transition-Metal–Hydrogen Complexes

Common transition metals possess electronic states belonging to the 3d, 4d or 5d shell. As such states do not exist for silicon atoms or crystals, entirely new hybrid electronic structures can be created. Transition-metal impurities, therefore, unlike shallow dopants, form a strong perturbation in the electronic structure of silicon, creating deep potentials that provide options for binding electrons or holes in deep bandgap levels. Trapping and releasing carriers via these states is a relatively fast process, rendering the transition metals active recombination centres, governing carrier lifetimes. For hydrogen also, and for several other impurities as well, the transition metals are strongly attractive centres, with the result that impurity complexes are abundant. The high diffusivity of, especially, the later transition elements in the 3d, 4d and 5d series (iron, copper, silver and gold), together with their natural presence adds to the reality of such defect formation processes, either intentional or as the result of insufficient control over contamination in the environment.

In recent years extensive ab-initio theoretical calculations have been performed on the existence of transition-metal–hydrogen complexes, notably for gold, silver, palladium and platinum [179–185]. Using spin-polarised wave

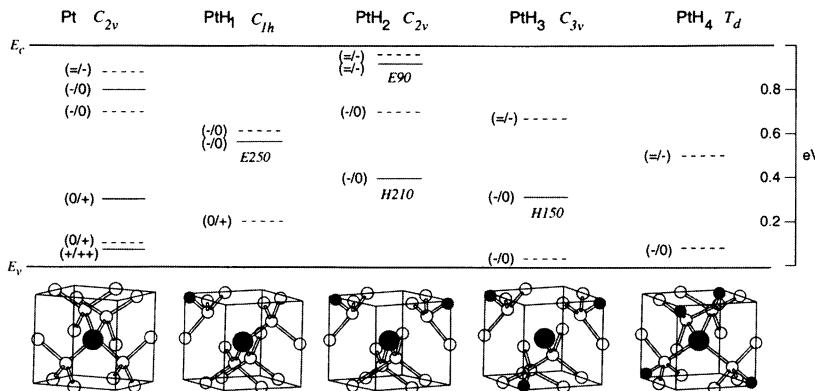


Fig. 14.11. Structure and corresponding electrical levels of platinum and its complexes with hydrogen in silicon, from experiment [199, 204, 205] and theory [182]. Experimental and calculated levels are represented by *solid* and *dashed lines*, respectively. After Jones et al. [185]

functions and the local-density formalism, the stability and electronic levels of transition metals in clusters containing over 100 silicon atoms were calculated. It was found that these transition elements can bind from one to four hydrogen atoms, creating new electrically active centres of donor, (multiple) acceptor or amphoteric character. Unlike the shallow dopants, the transition elements are not passivated. For platinum, the calculated electronic levels of the isolated impurity and its hydrides PtH_n , with $n = 1, 2, 3$ and 4, are indicated in Fig. 14.11 [185].

Extensive experimental investigations of the electronic structure have also been carried out using DLTS. Recent results for Pd, Ag, Pt and Au, which are not always in agreement with data obtained in earlier days [186–192], are included in Fig. 14.11 as well [193–207]. A special analysis based on depth profiles has allowed the determination of hydrogen atom numbers [208]. Though in general good agreement is claimed, several discrepancies are apparent. In the experiments, no levels were detected for the TM–H₄ complexes, leading to the conclusion that four hydrogen atoms passivate all these transition metals. Also, the coincidence of calculated and experimentally measured levels is far from satisfactory at present, asking for further upgrading. Another interesting feature is provided by the different donor and acceptor level positions for gold and silver in the theory, whereas in the experiments they are reported to be equal within the error margin, handicapping the distinguishing of these impurities [185].

Experimental investigations of 3d transition metals have revealed the formation of several hydrogenated but still electrically active complexes, with up to nine new levels in the bandgap for cobalt, for example. Detailed studies by DLTS were carried out for titanium [191, 202, 209], vanadium [191, 210, 211],

chromium [191, 210–212], iron [189, 190], cobalt [202, 213–215], nickel [188, 189, 202, 216] and copper [188, 189, 217]. In addition, for the 4d element rhodium, the hydrogen complexes RhH_1 and RhH_2 were reported, both with two levels [218]. Complexes dissociate, recreating the isolated transition metals, at temperatures typically below 300°C. Information on the atomic and electronic structure from experiment has been provided by observations of local vibrational modes in optical absorption and from the Zeeman effect and hyperfine interactions in magnetic resonance.

By far the most thorough investigations were performed on the Pt- and Au-related complexes, in particular on the centre identified as PtH_2 [52, 219–222]. This centre can be produced as a bulk defect, allowing IR and EPR measurements to be made, by hydrogenation treatment for 24 to 72 hours at 1000 to 1250°C of silicon doped with platinum. The magnetic resonance spectrum was analysed with the electron spin $S = 1/2$ and shows the angular dependence of a centre with the orthorhombic-I symmetry. Atomic constituents were identified on the basis of the observed hyperfine splitting patterns. The presence of one platinum atom was demonstrated by resolved hyperfine structure in three components with an intensity ratio of 0.25 : 1 : 0.25, as expected for natural platinum, where the isotope ^{195}Pt has a nuclear spin $I = 1/2$ and 33% abundance. Two hydrogen atoms (nuclear spin $I = 1/2$, abundance 100%) on symmetry-equivalent positions were revealed by hyperfine structure with a ratio 1 : 2 : 1 in each of the platinum lines. Such an EPR spectrum is shown in Fig. 14.12 [223]. If the hydrogen is replaced by deuterium, spin $I = 1$, the number of hyperfine lines will increase to five, with

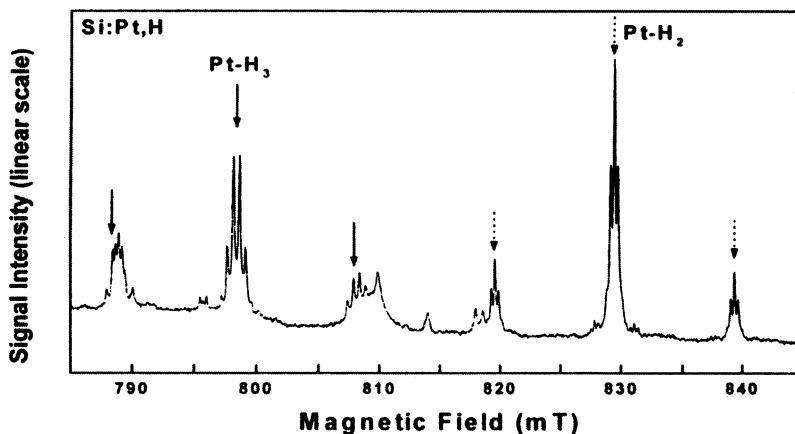


Fig. 14.12. EPR spectra of the centres Si:Pth_2 and Si:Pth_3 . Hyperfine interactions with the platinum isotopes and with hydrogen characterise the structure of the spectra, indicated by *dashed arrows* for PtH_2 and *solid arrows* for PtH_3 . After Huy and Ammerlaan [223]

an intensity ratio of 1:2:3:2:1, but their mutual separation is too small to be resolved and only the appropriate line broadening is observed. Local vibrational modes at 1888.2 cm^{-1} and 1901.6 cm^{-1} were identified as the anti-symmetric and symmetric hydrogen stretching-mode vibrations, respectively, of PtH_2 in its paramagnetic state [222]. The observation of isotope shifts in the vibration frequencies for the centres PtHD and PtD_2 confirms the presence of two hydrogen atoms in the centre [224]. Unlike EPR, vibrational spectroscopy is not restricted to the paramagnetic state $(\text{PtH}_2)^-$ of the centre and hence the corresponding vibrations for two other diamagnetic charge states of the centre have also been reported [224, 225]. The simultaneous loss of both LVM and EPR spectra upon thermal annealing at 600°C indicates the association of the spectra with the same centre [219]. Careful investigations were made of the charge state of the centre by monitoring the presence of spectra as the Fermi level was varied. As a result it was concluded that the paramagnetic state of the complex corresponds to $(\text{PtH}_2)^-$ and that a level $(\text{PtH}_2)^{2-}/(\text{PtH}_2)^-$ lies between $E_c - 0.045\text{ eV}$ and $E_c - 0.1\text{ eV}$. The first acceptor level $(\text{PtH}_2)^-/(\text{PtH}_2)^0$ is positioned between $E_c - 0.23\text{ eV}$ and $E_v + 0.32\text{ eV}$. These results are consistent with the DLTS data presented in Fig. 14.11. It must be concluded that hydrogenation of platinum-doped silicon leads to the formation of a double acceptor, and hence no electrical passivation. In the defect model the two hydrogen atoms have an interstitial position in an (011) plane through the substitutional platinum atom. In the analysis of the data a preference was deduced for silicon antibonding positions, outside the nearest silicon neighbours of the Pt atom. The position inside the nearest-neighbour cage, more of BC character, however, cannot be excluded yet.

EPR observations were also made of the PtH_3 complex, with trigonal symmetry [52, 223, 226, 227]. The presence of three symmetry-equivalent hydrogen atoms is deduced from the characteristic hyperfine splitting into four components with intensities in the ratio 1:3:3:1 as observed for hydrogen, and as illustrated in Fig. 14.12. Another similar centre is $\text{Au}_s(\text{H}_i)_2$, the gold analogue of PtH_2 , identified by the EPR spectrum Si-NL64 [52, 223, 226–228]. For this centre, the presence of one gold atom is indicated by four equal-amplitude resonances, reflecting the ^{197}Au isotope with its 100% abundance and nuclear spin $I = 3/2$. Figure 14.13 [228] shows a recorded resonance with the combined gold and hydrogen hyperfine interactions leading to the structure (1:2:1):(1:2:1):(1:2:1):(1:2:1). In LVM spectroscopy, the monohydride complexes PtH_1 and AuH_1 were identified as electrically active defects, in agreement with findings from theory and DLTS [224, 229, 230]. Remarkably, EPR spectra corresponding to these complexes have escaped detection for as yet unknown reasons. For all three EPR centres discussed, the isotropic part of the hydrogen hyperfine interaction is near 10 MHz. On comparing this strength with the coupling of an electron in the 1s state of hydrogen, with $a \approx 1400\text{ MHz}$, one concludes that spin density on hydrogen in the complexes

is very low. This can correspond to no electron in the 1s orbital and a positive hydrogen ion, or to a full 1s shell with two electrons of opposite spin and hydrogen as a negative ion. It appears that the neutral charge state H^0 is avoided, and one is tempted to conclude that the negative- U effect for isolated hydrogen, rendering H^0 unstable, is active in TM–H complexes as well.

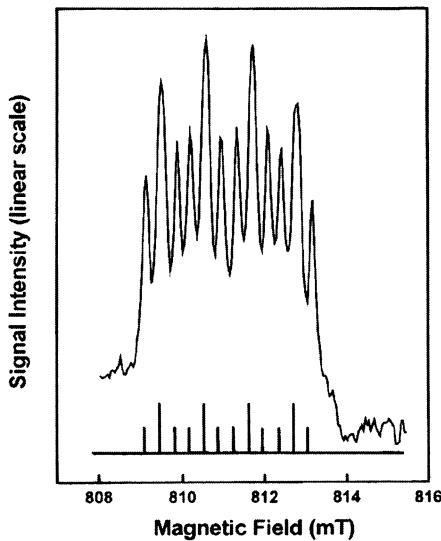


Fig. 14.13. Part of the EPR spectrum Si-NL64 of the AuH_2 centre in silicon. The fourfold splitting is due to hyperfine interaction with the ^{197}Au isotope (abundance 100%, nuclear spin $I = 3/2$), the additional 1:2:1 structure reveals the presence of two hydrogen atoms. After Huy and Ammerlaan [228]

14.6 Conclusion

This review has demonstrated the activity of the hydrogen impurity in silicon for a few important cases. The treatment of the subject has been far from exhaustive. Some topics not dealt with are the interactions of hydrogen with the intrinsic crystal defects (multi)vacancies and interstitials, produced by radiation and implantation; the effect opposite to passivation manifested by the activation by hydrogen of isoelectronic impurities, such as C; the phenomenon of hydrogen-enhanced diffusion of impurities, as observed for aluminium and oxygen; the formation of hydrogen-related shallow donors; the impact of hydrogen at the Si/SiO_2 interface, at grain boundaries and its interaction with P_b centres; and hydrogen in amorphous silicon, with its application in improving solar cell materials. For these subjects, reference should be made to the literature.

References

1. S.J. Pearton, J.W. Corbett, T.S. Shi: *Appl. Phys. A* **43**, 153 (1987)
2. J.I. Pankove, N.M. Johnson (eds.): *Hydrogen in Semiconductors*, Semicond. Semimetals **34** (1991)
3. M. Stutzmann, J. Chevallier (eds.): *Hydrogen in Semiconductors: Bulk and Surface Properties*, *Physica B* **170**, 1–581 (1991)
4. S.M. Myers, M.I. Baskes, H.K. Birnbaum, J.W. Corbett, G.G. DeLeo, S.K. Estreicher, E.E. Haller, P. Jena, N.M. Johnson, R. Kirchheim, S.J. Pearton, M.J. Stavola: *Rev. Mod. Phys.* **64**, 559 (1992)
5. S.J. Pearton, J.W. Corbett, M. Stavola: *Hydrogen in Crystalline Semiconductors* (Springer, Berlin 1992) pp. 1–363
6. S.K. Estreicher: *Mater. Sci. Eng. Rep. R* **14**, 317 (1995)
7. N.H. Nickel, W.B. Jackson, R.C. Bowman, R.G. Leisure (eds.): *Hydrogen in Semiconductors and Metals* (Materials Research Society, Warrendale 1998) pp. 1–455
8. N.H. Nickel (ed.): *Hydrogen in Semiconductors II*, Semicond. Semimetals **61** (1999)
9. M. Stavola: Hydrogen diffusion and solubility in c-Si. In: *Properties of Crystalline Silicon*, EMIS Datareviews Series No. 20, ed. by R. Hull (INSPEC, IEE, London 1999) pp. 511–521
10. M. Stavola: Hydrogen-containing point defects in c-Si. In: *Properties of Crystalline Silicon*, EMIS Datareviews Series No. 20, ed. by R. Hull (INSPEC, IEE, London 1999) pp. 522–537
11. J. Weber, A. Mesli (eds.): *Defects in Silicon: Hydrogen*, Mater. Sci. Eng. B **58**, 1–183 (1999)
12. R. Jones, B.J. Coomer, J.P. Goss, B. Hourahine, A. Resende: *Solid State Phenom.* **71**, 173 (2000)
13. J. Chevallier, B. Pajot: *Solid State Phenom.* **85–86**, 203 (2002)
14. C.A.J. Ammerlaan, P.T. Huy: *Solid State Phenom.* **85–86**, 353 (2002)
15. S. Estreicher: *Phys. Rev. B* **36**, 9122 (1987)
16. B. Bech Nielsen: *Phys. Rev. B* **37**, 6353 (1988)
17. P. Deák, L.C. Snyder, J.L. Lindström, J.W. Corbett, S.J. Pearton, A.J. Tavendale: *Phys. Lett. A* **126**, 427 (1988)
18. P. Deák, L.C. Snyder, J.W. Corbett: *Phys. Rev. B* **37**, 6887 (1988)
19. G.G. DeLeo, M.J. Dorogi, W.B. Fowler: *Phys. Rev. B* **38**, 7520 (1988)
20. C.G. Van de Walle, Y. Bar-Yam, S.T. Pantelides: *Phys. Rev. Lett.* **60**, 2761 (1988)
21. K.J. Chang, D.J. Chadi: *Phys. Rev. B* **40**, 11644 (1989)
22. C.G. Van de Walle, P.J.H. Denteneer, Y. Bar-Yam, S.T. Pantelides: *Phys. Rev. B* **39**, 10791 (1989)
23. C.G. Van de Walle: *Phys. Rev. B* **49**, 4579 (1994)
24. A.J. Tavendale, S.J. Pearton, A.A. Williams: *Appl. Phys. Lett.* **56**, 949 (1990)
25. J. Zhu, N.M. Johnson, C. Herring: *Phys. Rev. B* **41**, 12354 (1990)
26. N.M. Johnson, C. Herring: *Phys. Rev. B* **46**, 15554 (1992)
27. N.M. Johnson, C.G. Van de Walle: Isolated Monatomic Hydrogen in Silicon. *Semicond. Semimetals* **61**, 13 (1999)
28. C.H. Chu, S.K. Estreicher: *Phys. Rev. B* **42**, 9486 (1990)
29. C.G. Van de Walle: *Physica B* **170**, 21 (1991)

30. Yu.V. Gorelkinskii, N.N. Nevinnyi: Mater. Sci. Eng. B **36**, 133 (1996)
31. K. Irmscher, H. Klose, K. Maass: J. Phys. C: Solid State Phys. **17**, 6317 (1984)
32. B. Holm, K. Bonde Nielsen, B. Bech Nielsen: Phys. Rev. Lett. **66**, 2360 (1991)
33. B. Bech Nielsen, K. Bonde Nielsen, J.R. Byberg: Mater. Sci. Forum **143–147**, 909 (1994)
34. L.C. Kimerling, P. Blood, W.M. Gibson: Defect states in proton-bombarded silicon at $T < 300$ K. In: *Defects and Radiation Effects in Semiconductors 1978*, ed. by J.H. Albany (Institute of Physics, Bristol 1979) pp. 273–280
35. N.M. Johnson, C. Herring, C.G. Van de Walle: Phys. Rev. Lett. **73**, 130 (1994)
36. N.M. Johnson, C. Herring, C.G. Van de Walle: Phys. Rev. Lett. **74**, 4566 (1995)
37. C.H. Seager, R.A. Anderson, S.K. Streicher: Phys. Rev. Lett. **74**, 4565 (1995)
38. P. Deák, L.C. Snyder, J.W. Corbett: Phys. Rev. B **43**, 4545 (1991)
39. Yu.V. Gorelkinskii, N.N. Nevinnyi: Sov. Tech. Phys. Lett. **13**, 45 (1987) [Pis'ma Zh. Tekh. Fiz. (USSR) **13**, 105 (1987)]
40. Yu.V. Gorelkinskii, N.N. Nevinnyi: Physica B **170**, 155 (1991)
41. Yu.V. Gorelkinskii: Electron Parametric Resonance Studies of Hydrogen and Hydrogen-Related Defects in Crystalline Silicon. Semicond. Semimetals **61**, 25 (1999)
42. V.A. Gordeev, Yu.V. Gorelkinskii, R.F. Konopleva, N.N. Nevinnyi, Yu.V. Obukhov, V.G. Firsov: Preprint 1340, Leningrad Nuclear Physics Institute, Academy of Sciences of the USSR (1987), pp. 1–30
43. R.F. Kiefl, M. Celio, T.L. Estle, S.R. Kreitzman, G.M. Luke, T.M. Riseman, E.J. Ansaldi: Phys. Rev. Lett. **60**, 224 (1988)
44. R.B. Gel'fand, V.A. Gordeev, Yu.V. Gorelkinskii, R.F. Konopleva, S.A. Kuten', A.V. Mudryi, N.N. Nevinnyi, Yu.V. Obukhov, V.I. Rapoport, A.G. Ul'yashin, V.G. Firsov: Sov. Phys. Solid State **31**, 1376 (1989) [Fiz. Tverd. Tela **31**, 176 (1989)]
45. V.A. Gordeev, R.F. Konopleva, V.G. Firsov, Yu.V. Obukhov, Yu.V. Gorelkinskii, N.N. Nevinnyi: Hyperfine Interact. **60**, 717 (1990)
46. B. Hitti, S.R. Kreitzman, T.L. Estle, E.S. Bates, M.R. Dawdy, T.L. Head, R.L. Lichti: Phys. Rev. B **59**, 4918 (1999)
47. T.L. Estle, S. Streicher, D.S. Marynick: Phys. Rev. Lett. **58**, 1547 (1987)
48. B.D. Patterson, A. Hintermann, W. Kündig, P.F. Meier, F. Waldner, H. Graf, E. Recknagel, A. Weidinger, T. Wichert: Phys. Rev. Lett. **40**, 1347 (1978)
49. K.W. Blazey, J.A. Brown, D.W. Cooke, S.A. Dodds, T.L. Estle, R.H. Heffner, M. Leon, D.A. Vanderwater: Phys. Rev. B **23**, 5316 (1981)
50. K.W. Blazey, T.L. Estle, E. Holzschuh, W. Odermatt, B.D. Patterson: Phys. Rev. B **27**, 15 (1983)
51. C.G. Van de Walle, P.E. Blöchl: Phys. Rev. B **47**, 4244 (1993)
52. C.A.J. Ammerlaan, P.T. Huy: Solid State Phenom. **85–86**, 353 (2002)
53. A. van Wieringen, N. Warmoltz: Physica **22**, 849 (1956)
54. C. Herring, N.M. Johnson, C.G. Van de Walle: Phys. Rev. B **64**, 125209-1 (2001)
55. F. Buda, G.L. Chiarotti, R. Car, M. Parrinello: Phys. Rev. Lett. **63**, 294 (1989)
56. P.E. Blöchl, C.G. Van de Walle, S.T. Pantelides: Phys. Rev. Lett. **64**, 1401 (1990)

57. C. Herring, N.M. Johnson: Hydrogen Migration and Solubility in Silicon. *Semicond. Semimetals* **34**, 225 (1991)
58. N.M. Johnson, C. Herring: *Phys. Rev. B* **38**, 1581 (1988)
59. J.D. Holbech, B. Bech Nielsen, R. Jones, P. Sitch, S. Öberg: *Phys. Rev. Lett.* **71**, 875 (1993)
60. K.J. Chang, D.J. Chadi: *Phys. Rev. Lett.* **62**, 937 (1989)
61. S.B. Zhang, W.B. Jackson, D.J. Chadi: *Phys. Rev. Lett.* **65**, 2575 (1990)
62. A.N. Safonov, E.C. Lightowers, G. Davies: *Phys. Rev. B* **56**, R15517 (1997)
63. B.N. Mukashev, K.H. Nussupov, M.F. Tamendarov: *Phys. Lett.* **72A**, 381 (1979)
64. H.J. Stein: *Phys. Rev. Lett.* **43**, 1030 (1979)
65. B.N. Mukashev, K.H. Nussupov, M.F. Tamendarov, V.V. Frolov: *Phys. Lett.* **87A**, 376 (1982)
66. K.J. Chang, D.J. Chadi: *Phys. Rev. B* **42**, 7651 (1990)
67. J.I. Pankove, D.E. Carlson, J.E. Berkeyheiser, R.O. Wance: *Phys. Rev. Lett.* **51**, 2224 (1983)
68. C.-T. Sah, J.Y.-C. Sun, J.J.-T. Tzou: *Appl. Phys. Lett.* **43**, 204 (1983)
69. J.I. Pankove, R.O. Wance, J.E. Berkeyheiser: *Appl. Phys. Lett.* **45**, 1100 (1984)
70. Y.-C. Du, Y.-F. Zhang, G.-G. Qin, S.-F. Weng: *Solid State Commun.* **55**, 501 (1985)
71. A. Chari, M. Aucouturier: *Solid State Commun.* **71**, 105 (1989)
72. M.L.W. Thewalt, E.C. Lightowers, J.I. Pankove: *Appl. Phys. Lett.* **46**, 689 (1985)
73. G.G. DeLeo, W.B. Fowler: *Phys. Rev. B* **31**, 6861 (1985)
74. N.M. Johnson: *Phys. Rev. B* **31**, 5525 (1985)
75. J.I. Pankove, D.J. Zanzucchi, C.W. Magee, G. Lucovsky: *Appl. Phys. Lett.* **46**, 421 (1985)
76. M. Stutzmann: *Phys. Rev. B* **35**, 5921 (1987)
77. M. Stavola, S.J. Pearton, J. Lopata, W.C. Dautremont-Smith: *Appl. Phys. Lett.* **50**, 1086 (1987)
78. B. Pajot, A. Chari, M. Aucouturier, M. Astier, A. Chantre: *Solid State Commun.* **67**, 855 (1988)
79. G.G. DeLeo, W.B. Fowler: *J. Electron. Mater.* **14a**, 745 (1985)
80. G.G. DeLeo, W.B. Fowler: *Phys. Rev. Lett.* **56**, 402 (1986)
81. A. Amore Bonapasta, A. Lapicciarella, N. Tomassini, M. Capizzi: *Phys. Rev. B* **36**, 6228 (1987)
82. A.D. Marwick, G.S. Oehrlein, N.M. Johnson: *Phys. Rev. B* **36**, 4539 (1987)
83. B. Bech Nielsen, J.U. Andersen, S.J. Pearton: *Phys. Rev. Lett.* **60**, 321 (1988)
84. K. Bergman, M. Stavola, S.J. Pearton, T. Hayes: *Phys. Rev. B* **38**, 9643 (1988)
85. K.J. Chang, D.J. Chadi: *Phys. Rev. Lett.* **60**, 1422 (1988)
86. A.D. Marwick, G.S. Oehrlein, J.H. Barrett, N.M. Johnson: The structure of the boron-hydrogen complex in silicon. In: *Defects in Electronic Materials*, ed. by M. Stavola, S.J. Pearton, G. Davies (Materials Research Society, Pittsburgh 1988) pp. 259–264
87. M. Stavola, K. Bergman, S.J. Pearton, J. Lopata: *Phys. Rev. Lett.* **61**, 2786 (1988)

88. C.G. Van de Walle, P.J.H. Denteneer, Y. Bar-Yam, S.T. Pantelides: Hydrogen diffusion and passivation of shallow impurities in crystalline silicon. In: *Shallow Impurities in Semiconductors 1988*, ed. by B. Monemar (Institute of Physics, Bristol 1989) pp. 405–414
89. E. Artacho, F. Ynduráin: Solid State Commun. **72**, 393 (1989)
90. P.J.H. Denteneer, C.G. Van de Walle, S.T. Pantelides: Phys. Rev. B **39**, 10809 (1989)
91. P.J.H. Denteneer, C.G. Van de Walle, S.T. Pantelides: Phys. Rev. Lett. **62**, 1884 (1989)
92. P.J.H. Denteneer, C.G. Van de Walle, Y. Bar-Yam, S.T. Pantelides: Mater. Sci. Forum **38–41**, 979 (1989)
93. S.K. Estreicher, L. Throckmorton, D.S. Marynick: Phys. Rev. B **39**, 13241 (1989)
94. T. Sasaki, H. Katayama-Yoshida: Mechanism of hydrogen passivation in silicon. In: *Shallow Impurities in Semiconductors 1988*, ed. by B. Monemar (Institute of Physics, Bristol 1989) pp. 395–404
95. T. Zundel, J. Weber: Phys. Rev. B **39**, 13549 (1989)
96. G.D. Watkins, W.B. Fowler, M. Stavola, G.G. DeLeo, D.M. Kozuch, S.J. Pearton, J. Lopata: Phys. Rev. Lett. **64**, 467 (1990)
97. K.R. Martin, W.B. Fowler, G.G. DeLeo: Mater. Sci. Forum **83–87**, 69 (1992)
98. Dj.M. Maric, P.F. Meier, S.K. Estreicher: Phys. Rev. B **47**, 3620 (1993)
99. Y. Zhou, R. Luchsinger, P.F. Meier: Phys. Rev. B **51**, 4166 (1995)
100. G.G. DeLeo, W.B. Fowler: Computational Studies of Hydrogen-Containing Complexes in Semiconductors. Semicond. Semimetals **34**, 511 (1991)
101. Dj.M. Maric, P.F. Meier: Helv. Phys. Acta **64**, 908 (1991)
102. C.P. Herrero, M. Stutzmann, A. Breitschwerdt: Phys. Rev. B **43**, 1555 (1991)
103. C.P. Herrero, M. Stutzmann: Solid State Commun. **68**, 1085 (1988)
104. M. Stutzmann, C.P. Herrero: Raman studies of hydrogen passivation in silicon. In: *Defects in Electronic Materials*, ed. by M. Stavola, S.J. Pearton, G. Davis (Materials Research Society, Pittsburgh 1988) pp. 271–276
105. A. Amore Bonapasta, P. Giannozzi, M. Capizzi: Phys. Rev. B **44**, 3399 (1991)
106. A. Amore Bonapasta, P. Giannozzi, M. Capizzi: Phys. Rev. B **45**, 11744 (1992)
107. Y. Zhou, R. Luchsinger, P.F. Meier: Mater. Sci. Forum **196–201**, 885 (1995)
108. M. Stavola, S.J. Pearton, J. Lopata, W.C. Dautremont-Smith: Phys. Rev. B **37**, 8313 (1988)
109. L.V.C. Assali, J.R. Leite: Phys. Rev. Lett. **55**, 980 (1985)
110. L.V.C. Assali, J.R. Leite: Phys. Rev. Lett. **56**, 403 (1986)
111. E.C.F. da Silva, L.V.C. Assali, J.R. Leite, A. Dal Pino Jr.: Phys. Rev. B **37**, 3113 (1988)
112. A. Baurichter, S. Deubler, D. Forkel, M. Uhrmacher, H. Wolf, W. Witthuhn: Hydrogen passivation of indium acceptors in silicon. In: *Shallow Impurities in Semiconductors 1988*, ed. by B. Monemar (Institute of Physics, Bristol 1989) pp. 471–476

113. T. Wichert, H. Skudlik, M. Deicher, G. Grübel, R. Keller, E. Recknagel, L. Song: *Phys. Rev. Lett.* **59**, 2087 (1987)
114. T. Wichert, H. Skudlik, H.-D. Carstanjen, T. Enders, M. Deicher, G. Grübel, R. Keller, L. Song, M. Stutzmann: Localization of hydrogen in B and In doped silicon by ion channeling and PAC. In: *Defects in Electronic Materials*, ed. by M. Stavola, S.J. Pearton, G. Davies (Materials Research Society, Pittsburgh 1988) pp. 265–270
115. J.M. Baranowski, J. Tatarkiewicz: *Phys. Rev. B* **35**, 7450 (1987)
116. N.M. Johnson, C. Doland, F. Ponce, J. Walker, G. Anderson: *Physica B* **170**, 3 (1991)
117. M. Stavola, K. Bergman, S.J. Pearton, J. Lopata, T. Hayes: The symmetry and properties of donor-H and acceptor-H complexes in Si from uniaxial stress studies. In: *Shallow Impurities in Semiconductors 1988*, ed. by B. Monemar (Institute of Physics, Bristol 1989) pp. 447–452
118. G. Cannelli, R. Cantelli, M. Capizzi, C. Coluzza, F. Cordero, A. Frova, A. Lo Presti: *Phys. Rev. B* **44**, 11486 (1991)
119. M. Stavola, Y.M. Cheng: *Solid State Commun.* **93**, 431 (1995)
120. M. Stavola, J.-F. Zheng, Y.M. Cheng, C.R. Abernathy, S.J. Pearton: *Mater. Sci. Forum* **196–201**, 809 (1995)
121. M. Stavola, Y.M. Cheng, G. Davies: *Mater. Sci. Forum* **143–147**, 885 (1994)
122. C.H. Seager, R.A. Anderson: *Appl. Phys. Lett.* **59**, 585 (1991)
123. T. Zundel, J. Weber: *Phys. Rev. B* **43**, 4361 (1991)
124. T. Zundel, J. Weber, L. Tilly: *Physica B* **170**, 361 (1991)
125. Y. Ohmura, K. Abe, M. Ohtaka, A. Kimoto, M. Yamaura: *Mater. Sci. Forum* **258–263**, 185 (1997)
126. K. Muro, A.J. Sievers: *Phys. Rev. Lett.* **57**, 897 (1986)
127. E.E. Haller: Hydrogen-related effects in crystalline semiconductors. In: *Shallow Impurities in Semiconductors 1988*, ed. by B. Monemar (Institute of Physics, Bristol 1989) pp. 425–436
128. R.E. Peale, K. Muro, A.J. Sievers: *Phys. Rev. B* **41**, 5881 (1990)
129. M. Gebhard, B. Vogt, W. Witthuhn: *Phys. Rev. Lett.* **67**, 847 (1991)
130. M. Suezawa, R. Mori: *Phys. Stat. Sol. (b)* **210**, 507 (1998)
131. M. Suezawa: Optical absorption study of hydrogen in Zn-doped Si. In: *Hydrogen in Semiconductors and Metals*, ed. by N.H. Nickel, W.B. Jackson, R.C. Bowman, R.G. Leisure (Materials Research Society, Warrendale 1998) pp. 357–362
132. R. Mori, M. Suezawa: *Physica B* **273–274**, 220 (1999)
133. R. Mori, N. Fukata, M. Suezawa, A. Kasuya: *Physica B* **302–303**, 206 (2001)
134. P. Stolz, G. Pensl, D. Grünebaum, N. Stolwijk: *Mater. Sci. Eng. B* **4**, 31 (1989)
135. K. Bergman, M. Stavola, S.J. Pearton, J. Lopata: *Phys. Rev. B* **37**, 2770 (1988)
136. Z.N. Liang, L. Niesen: *Nucl. Instrum. Methods Phys. Res. B* **63**, 147 (1992)
137. N.M. Johnson, C. Herring, D.J. Chadi: *Phys. Rev. Lett.* **56**, 769 (1986)
138. A. Amore Bonapasta, P. Giannozzi, M. Capizzi: *Phys. Rev. B* **42**, 3175 (1990)
139. P.J.H. Denteneer, C.G. Van de Walle, S.T. Pantelides: *Phys. Rev. B* **41**, 3885 (1990)
140. S.B. Zhang, D.J. Chadi: *Phys. Rev. B* **41**, 3882 (1990)
141. Z.N. Liang, P.J.H. Denteneer, L. Niesen: *Phys. Rev. B* **52**, 8864 (1995)

142. A. Amore Bonapasta, A. Lapiccirella, N. Tomassini, M. Capizzi: Phys. Rev. B **39**, 12630 (1989)
143. G.G. DeLeo, W.B. Fowler, T.M. Sudol, K.J. O'Brien: Phys. Rev. B **41**, 7581 (1990)
144. N.M. Johnson, C. Herring: Hydrogen neutralization of shallow-donor impurities in single-crystal silicon. In: *Defects in Electronic Materials*, ed. by M. Stavola, S.J. Pearton, G. Davies (Materials Research Society, Pittsburgh 1988), pp. 277–280
145. Z.N. Liang, C. Haas, L. Niesen: Phys. Rev. Lett. **72**, 1846 (1994)
146. Z.N. Liang, L. Niesen: Phys. Rev. B **51**, 11120 (1995)
147. L. Korpás, J.W. Corbett, S.K. Estreicher: Mater. Sci. Forum **83–87**, 27 (1992)
148. L. Korpás, J.W. Corbett, S.K. Estreicher: Phys. Rev. B **46**, 12365 (1992)
149. S.P. Love, K. Muro, R.E. Peale, A.J. Sievers, W. Lo: Phys. Rev. B **36**, 2950 (1987)
150. R.E. Peale, K. Muro, A.J. Sievers: Mater. Sci. Forum **65–66**, 151 (1990)
151. I.S. Zevenbergen, T. Gregorkiewicz, C.A.J. Ammerlaan: Phys. Rev. B **51**, 16746 (1995)
152. I.S. Zevenbergen, T. Gregorkiewicz, C.A.J. Ammerlaan: Mater. Sci. Forum **196–201**, 855 (1995)
153. C.A.J. Ammerlaan, I.S. Zevenbergen, T. Gregorkiewicz: Passivation of electronic centres in silicon by hydrogen. In: *Physics of Semiconductor Devices*, ed. by V. Kumar, S.K. Agarwal (Narosa, New Delhi 1998) pp. 531–538
154. C.A.J. Ammerlaan, P.T. Huy: Solid State Phenom. **69–70**, 583 (1999)
155. P.T. Huy, C.A.J. Ammerlaan, T. Gregorkiewicz: Physica B **273–274**, 239 (1999)
156. P.T. Huy, C.A.J. Ammerlaan, T. Gregorkiewicz, D.T. Don: Phys. Rev. B **61**, 7448 (2000)
157. G. Pensl, G. Roos, C. Holm, E. Sirtl, N.M. Johnson: Appl. Phys. Lett. **51**, 451 (1987)
158. G. Pensl, G. Roos, P. Stolz, N.M. Johnson, C. Holm: Hydrogen neutralization of chalcogen double donor centers in single-crystal silicon. In: *Defects in Electronic Materials*, ed. by M. Stavola, S.J. Pearton, G. Davies (Materials Research Society, Pittsburgh 1988) pp. 241–246
159. G. Roos, G. Pensl, N.M. Johnson, C. Holm: J. Appl. Phys. **67**, 1897 (1990)
160. Yu.V. Martynov, I.S. Zevenbergen, T. Gregorkiewicz, C.A.J. Ammerlaan: Solid State Phenom. **47–48**, 267 (1996)
161. I.S. Zevenbergen: Ph.D. thesis, University of Amsterdam (1998)
162. V.J.B. Torres, S. Öberg, R. Jones: Theory of hydrogen single passivated substitutional sulphur double donor in Si. In: *Shallow-Level Centers in Semiconductors*, ed. by C.A.J. Ammerlaan, B. Pajot (World Scientific, Singapore 1997) pp. 501–504
163. J. Coutinho, V.J.B. Torres, R. Jones, A. Resende, P.R. Briddon: Phys. Stat. Sol. (b) **235**, 107 (2003)
164. J. Coutinho, V.J.B. Torres, R. Jones, P.R. Briddon: Phys. Rev. B **67**, 035205-1 (2003)
165. A.S. Yapsir, P. Deák, R.K. Singh, L.C. Snyder, J.W. Corbett, T.-M. Lu: Phys. Rev. B **38**, 9936 (1988)
166. N.M. Johnson, S.K. Hahn, H.J. Stein: Mater. Sci. Forum **10–12**, 585 (1986)
167. N.M. Johnson, S.K. Hahn: Appl. Phys. Lett. **48**, 709 (1986)

168. A. Chantre, S.J. Pearton, L.C. Kimerling, K.D. Cummings, W.C. Dautremont-Smith: *Appl. Phys. Lett.* **50**, 513 (1987)
169. D.I. Bohne, J. Weber: *Phys. Rev. B* **47**, 4037 (1993)
170. D.I. Bohne, J. Weber: *Mater. Sci. Forum* **143–147**, 879 (1994)
171. J. Weber, D.I. Bohne: Passivation of thermal donors by atomic hydrogen. In: *Early Stages of Oxygen Precipitation in Silicon*, ed. by R. Jones (Kluwer Academic, Dordrecht 1996) pp. 123–140
172. Yu.V. Martynov, T. Gregorkiewicz, C.A.J. Ammerlaan: *Phys. Rev. Lett.* **74**, 2030 (1995)
173. Yu.V. Martynov, T. Gregorkiewicz, C.A.J. Ammerlaan: *Mater. Sci. Forum* **196–201**, 849 (1995)
174. C.A.J. Ammerlaan, I.S. Zevenbergen, Yu.V. Martynov, T. Gregorkiewicz: Magnetic resonance investigations of thermal donors in silicon. In: *Early Stages of Oxygen Precipitation in Silicon*, ed. by R. Jones (Kluwer Academic, Dordrecht 1996) pp. 61–82
175. P. Deák, L.C. Snyder, J.W. Corbett: *Phys. Rev. B* **45**, 11612 (1992)
176. R.C. Newman, M.J. Ashwin, R.E. Pritchard, J.H. Tucker, E.C. Lightowers, T. Gregorkiewicz, I.S. Zevenbergen, C.A.J. Ammerlaan, R. Falster, M.J. Binns: *Mater. Sci. Forum* **258–263**, 379 (1997)
177. R.E. Pritchard, M.J. Ashwin, J.H. Tucker, R.C. Newman, E.C. Lightowers, T. Gregorkiewicz, I.S. Zevenbergen, C.A.J. Ammerlaan, R. Falster, M.J. Binns: *Semicond. Sci. Technol.* **12**, 1404 (1997)
178. R.C. Newman, J.H. Tucker, N.G. Semaltianos, E.C. Lightowers, T. Gregorkiewicz, I.S. Zevenbergen, C.A.J. Ammerlaan: *Phys. Rev. B* **54**, R6803 (1996)
179. D.E. Woon, D.S. Marynick, S.K. Estreicher: *Phys. Rev. B* **45**, 13383 (1992)
180. R. Jones, S. Öberg, J. Goss, P.R. Briddon, A. Resende: *Phys. Rev. Lett.* **75**, 2734 (1995)
181. A. Resende, J. Goss, P.R. Briddon, S. Öberg, R. Jones: *Mater. Sci. Forum* **258–263**, 295 (1997)
182. R. Jones, A. Resende, S. Öberg, P.R. Briddon: *Mater. Sci. Eng. B* **58**, 113 (1999)
183. A. Resende, R. Jones, S. Öberg, P.R. Briddon: *Mater. Sci. Eng. B* **58**, 146 (1999)
184. A. Resende, R. Jones, S. Öberg, P.R. Briddon: *Phys. Rev. Lett.* **82**, 2111 (1999)
185. R. Jones, B.J. Coomer, J.P. Goss, B. Hourahine, A. Resende: *Solid State Phenom.* **71**, 173 (2000)
186. S.J. Pearton, A.J. Tavendale: *Phys. Rev. B* **26**, 7105 (1982)
187. S.J. Pearton, E.E. Haller: *J. Appl. Phys.* **54**, 3613 (1983)
188. S.J. Pearton, A.J. Tavendale: *J. Appl. Phys.* **54**, 1375 (1983)
189. A.J. Tavendale, S.J. Pearton: *J. Phys. C: Solid State Phys.* **16**, 1665 (1983)
190. S.J. Pearton, A.J. Tavendale: *J. Phys. C: Solid State Phys.* **17**, 6701 (1984)
191. R. Singh, S.J. Fonash, A. Rohatgi: *Appl. Phys. Lett.* **49**, 800 (1986)
192. A. Mesli, E. Courcelle, T. Zundel, P. Siffert: *Phys. Rev. B* **36**, 8049 (1987)
193. E.Ö. Sveinbjörnsson, O. Engström: *Appl. Phys. Lett.* **61**, 2323 (1992)
194. H. Feichtinger, E. Sturm: *Mater. Sci. Forum* **143–147**, 111 (1994)
195. E.Ö. Sveinbjörnsson, G.I. Andersson, O. Engström: *Phys. Rev. B* **49**, 7801 (1994)

196. E.Ö. Sveinbjörnsson, O. Engström: Mater. Sci. Forum **143–147**, 821 (1994)
197. E.Ö. Sveinbjörnsson, O. Engström: Phys. Rev. B **52**, 4884 (1995)
198. J.A. Davidson, J.H. Evans: Semicond. Sci. Technol. **11**, 1704 (1996)
199. J.-U. Sachse, E.Ö. Sveinbjörnsson, W. Jost, J. Weber, H. Lemke: Phys. Rev. B **55**, 16176 (1997)
200. J.-U. Sachse, J. Weber, H. Lemke: Mater. Sci. Forum **258–263**, 307 (1997)
201. N. Yarykin, J.-U. Sachse, J. Weber, H. Lemke: Mater. Sci. Forum **258–263**, 301 (1997)
202. J. Weber: Electrical properties of transition metal hydrogen complexes in silicon. In: *Hydrogen in Semiconductors and Metals*, ed. by N.H. Nickel, W.B. Jackson, R.C. Bowman, R.G. Leisure (Materials Research Society, Warrendale 1998) pp. 345–356
203. L. Rubaldo, P. Deixler, I.D. Hawkins, J. Terry, D.K. Maude, J.-C. Portal, J.H. Evans-Freeman, L. Dobaczewski, A.R. Peaker: Mater. Sci. Eng. B **58**, 126 (1999)
204. J.-U. Sachse, J. Weber, E.Ö. Sveinbjörnsson: Phys. Rev. B **60**, 1474 (1999)
205. J.-U. Sachse, E.Ö. Sveinbjörnsson, N. Yarykin, J. Weber: Mater. Sci. Eng. B **58**, 134 (1999)
206. J. Weber, S. Knack, J.-U. Sachse: Physica B **273–274**, 429 (1999)
207. N. Yarykin, J.-U. Sachse, H. Lemke, J. Weber: Phys. Rev. B **59**, 5551 (1999)
208. O.V. Feklisova, N.A. Yarykin: Semicond. Sci. Technol. **12**, 742 (1997)
209. W. Jost, J. Weber: Phys. Rev. B **54**, R11038 (1996)
210. T. Sadoh, H. Nakashima, T. Tsurushima: J. Appl. Phys. **72**, 520 (1992)
211. T. Sadoh, M. Watanabe, H. Nakashima, T. Tsurushima: Mater. Sci. Forum **143–147**, 939 (1994)
212. T. Sadoh, M. Watanabe, H. Nakashima, T. Tsurushima: J. Appl. Phys. **75**, 3978 (1994)
213. W. Jost, J. Weber, H. Lemke: Mater. Sci. Forum **196–201**, 927 (1995)
214. W. Jost, J. Weber, H. Lemke: Semicond. Sci. Technol. **11**, 22 (1996)
215. W. Jost, J. Weber, H. Lemke: Semicond. Sci. Technol. **11**, 525 (1996)
216. M. Shiraishi, J.-U. Sachse, H. Lemke, J. Weber: Mater. Sci. Eng. B **58**, 130 (1999)
217. S. Knack, J. Weber, H. Lemke: Physica B **273–274**, 387 (1999)
218. S. Knack, J. Weber, H. Lemke: Mater. Sci. Eng. B **58**, 141 (1999)
219. P.M. Williams, G.D. Watkins, S. Uftring, M. Stavola: Phys. Rev. Lett. **70**, 3816 (1993)
220. M. Höhne, U. Juda, Yu.V. Martynov, T. Gregorkiewicz, C.A.J. Ammerlaan, L.S. Vlasenko: Phys. Rev. B **49**, 13423 (1994)
221. M. Höhne, U. Juda, Yu.V. Martynov, T. Gregorkiewicz, C.A.J. Ammerlaan, L.S. Vlasenko: Mater. Sci. Forum **143–147**, 1659 (1994)
222. P.M. Williams, G.D. Watkins, S. Uftring, M. Stavola: Mater. Sci. Forum **143–147**, 891 (1994)
223. P.T. Huy, C.A.J. Ammerlaan: Solid State Phenom. **82–84**, 133 (2002)
224. S.J. Uftring, M. Stavola, P.M. Williams, G.D. Watkins: Phys. Rev. B **51**, 9612 (1995)
225. M. Stavola, S.J. Uftring, M.J. Evans, P.M. Williams, G.D. Watkins: Spectroscopy of transition-metal-hydrogen complexes in silicon. In: *Defect and Impurity Engineered Semiconductors and Devices*, ed. by S. Ashok, J. Chevalier, I. Akasaki, N.M. Johnson, B.L. Sopori (Materials Research Society, Pittsburgh 1995) pp. 341–352

226. P.T. Huy, C.A.J. Ammerlaan: *Physica B* **302–303**, 233 (2001)
227. P.T. Huy, C.A.J. Ammerlaan: *Physica B* **308–310**, 408 (2001)
228. P.T. Huy, C.A.J. Ammerlaan: *Phys. Rev. B* **66**, 165219-1 (2002)
229. M.J. Evans, M.G. Gornstein, M. Stavola: Vibrational spectroscopy of gold hydrogen complexes in silicon. In: *Defects in Electronic Materials II*, ed. by J. Michel, T. Kennedy, K. Wada, K. Thonke (Materials Research Society, Pittsburg 1997) pp. 275–280
230. M.J. Evans, M. Stavola, M.G. Weinstein, S.J. Uftring: *Mater. Sci. Eng. B* **58**, 118 (1999)

Part VII

Devices

15 Power Semiconductor Devices

A. Porst

15.1 Introduction

15.1.1 History

The most important step in the history of semiconductor devices was the discovery of the transistor effect in a semiconductor material by Bardeen and Brattain [1] and Shockley [2] at the end of the 1940s. This invention stimulated the further development of power semiconductor devices. Efforts were undertaken to realize the effects that were predicted to occur when a p-type (acceptor-doped) and an n-type (donor-doped) region were placed close together in a semiconductor material.

In 1952 a diode structure was proposed and also realized as a “large area diode” by Hall [3]. With a blocking voltage of 200 V and a forward current capability of 35 A, the first power semiconductor device was fabricated.

At that time germanium was used as the semiconductor material because it was more easily available. But very soon, theoretical investigations proved silicon’s greater potential for operating at elevated temperature, resulting from the higher band gap in comparison with germanium.

The development of the most important silicon power devices is shown on a time scale in Fig. 15.1 together with the voltage and current ranges of those devices as manufactured. Between 1950 and 1970, only diodes and thyristors were used as power devices. In the early years the development of the semiconductor devices had to be done in parallel with the development and production of the silicon material itself. In the middle of the 1950s power silicon rectifiers (diodes) were produced at the Siemens laboratory in Preetzfeld, Germany, showing blocking voltages higher than 1000 V and a current capability of about 150 A.

The activities in the development of power semiconductors at this time were driven by the need to substitute the large-volume mercury valves used in the field of electrolysis by much smaller semiconductor devices.

Also, a substitution of thyratrons by semiconductor devices could be expected after the proposal of a four-layer device in 1956 [4], later called the thyristor, which was realized in the 1960s. A third terminal facilitates a turn-on of the device at any time.

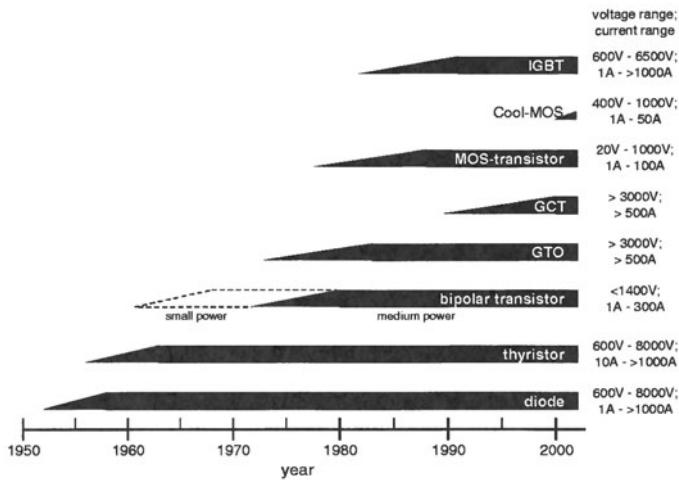


Fig. 15.1. History of power devices

In the 1970s the transistor for medium power and the GTO (gate turn-off thyristor) were introduced [5,6]. At this time, MOS transistors with a gate terminal isolated from the semiconductor surface by an oxide were used in integrated-circuit technology. MOS means a sequence of layers metal–oxide–semiconductor at the top of the chip. In integrated-circuit technology these transistors are designed as lateral structures, which are not suitable for high blocking voltages and high currents. But the dramatically reduced gate drive requirements in comparison with bipolar transistors intensified the attempts to realize power MOS transistors. At the end of the 1970s the first power MOS transistors existed containing a vertical structure with a current flow between the top and bottom of the chip [7,8]. One of the most popular power devices used today, the IGBT (insulated-gate bipolar transistor), is a modified MOS transistor with a bipolar conductivity, and it was presented in the early 1980s [9–12].

The MOS transistor is a majority-type device, with only one carrier type responsible for the behavior. In all other devices listed above, the existence of two types of carriers (electrons and holes) dominates their behavior. For these devices, the recombination between electrons and holes, the existence of field and diffusion currents of both types of carriers, and the possibility of neutral regions built up by the movable electrons and holes are the fundamental physical principles, which are described in [13], for example. The great advantages of semiconductor devices in comparison with mercury valves and thyratrons are much smaller volumes, distinctly lower losses, faster reaction times to control pulses and the fact that no wear has to be considered.

For a long time during the development of power devices (e.g. for diodes and thyristors) the size of the device was determined by the diameter of the

silicon rods produced by a technique that used a zone refining process, since one wafer was used for one device. Therefore the highest possible current capability was coupled to the maximum available diameter of the silicon rod.

When the bipolar transistor was introduced the paralleling of chips became necessary because the more failure-sensitive structures prevented an unlimited enlargement of the chip size. Such limitations are valid in a stronger way for MOS devices (MOS transistors and IGBT) with lateral patterns in the range of a few microns, where small defects in the various top layers (oxide, polysilicon and metal) may lower the yield dramatically. In order to conduct the current in and out and to protect the silicon devices against environmental risks, encapsulation is necessary. Large thyristors and GTOs are mounted into a disk cell (Fig. 15.2) and the electrical contacts on both sides are made by pressure, which has to be provided by the circuitry.

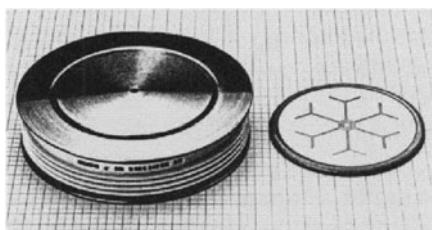


Fig. 15.2. 5500 V thyristor with a diameter of 100 mm (*right*), and encapsulated into a disk cell (*left*)

For chips with sizes below about 50 mm^2 , plastic cases are used. To be able to handle higher currents, chips (e.g. bipolar transistors or IGBTs) are arranged in parallel in a modular construction, which may involve the paralleling of 24 or 36 IGBT chips, each one of a size of about $14 \text{ mm} \times 14 \text{ mm}$ (Fig. 15.3). Today, current ratings up to 3600 A with a blocking-voltage capability of 1200 V are possible for such a device.

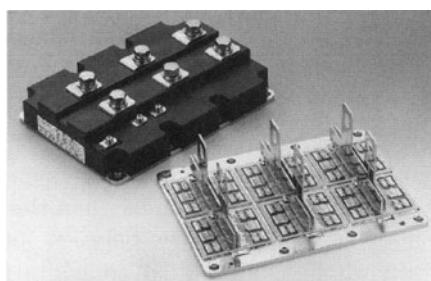


Fig. 15.3. Chips (24 IGBTs and 12 diodes) in a module arrangement

The chips are soldered onto an isolating copper-plated ceramic substrate on one side and are contacted by wires that are bonded onto the metallization of the chips on the other side. Additional chips (e.g. diodes) and arrangements of various switch configurations can also be included in such module constructions.

15.1.2 Requirements on Power Semiconductor Devices

Like mercury valves and thyratrons, semiconductor devices are used to convert and distribute electrical energy, including also the field of motor speed control, by performing the function of a switch for opening and closing an electrical circuit.

In the case of a diode – a two-terminal device – the function of a switch takes place automatically but depends on the polarity of the applied voltage. These devices are less suitable for active switching, but are indispensable in conjunction with active switches.

Devices with an additional gate terminal (three-terminal devices) make control possible. In the case of a thyristor, the turn-on can be controlled, but for the turn-off the polarity of the applied voltage has to be changed to reduce the current below a certain level. This change of polarity happens automatically in the case of an alternating current; otherwise, the reduction of the current has to be induced by auxiliary circuits.

In more modern three-terminal devices (e.g. the GTO, bipolar transistor, MOS transistor and IGBT) the turn-on and the turn-off can be controlled by the third terminal. Together with the progress in the development of the control circuits, these devices have made possible a wider field of application in motor speed control systems. Starting from the 50 Hz or 60 Hz line frequency, frequencies usually in the range between 0 and 200 Hz are generated using an intermediate direct-current circuit, e.g. a capacitor. With a variable frequency it is possible to control the speed of the cheap and widespread asynchronous motor in order to adapt it to the demands of the application.

By varying the time intervals for the on- and off-states of the switches S1 and S4 in the circuit shown in Fig. 15.4, a positive sine half-wave can be generated (Fig. 15.5), also with a variable frequency. The switches S2 and S3 are used for generating the negative sine half-wave.

If the switches in Fig. 15.4 are in the off-state, the diodes facilitate a continuous current flow in the load, preventing a current interruption with dangerous voltage spikes as a result.

In order to obtain a smooth sine wave, the frequency of the switching has to be at least an order of magnitude higher than the resulting sine wave frequency. Switching frequencies in the range between 500 Hz and 20 kHz are usual. By arranging six switches in a similar configuration and by suitable switching, three-phase sine wave currents can be generated.

In order to realize the functions of a switch, the properties of a power semiconductor device must include:

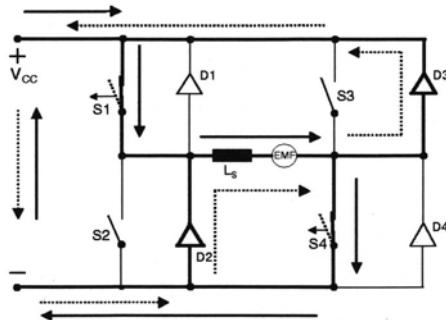


Fig. 15.4. Circuit used for the generation of a sine wave from a direct voltage source, and current flow with switches S1 and S4 closed (—) and open (···)

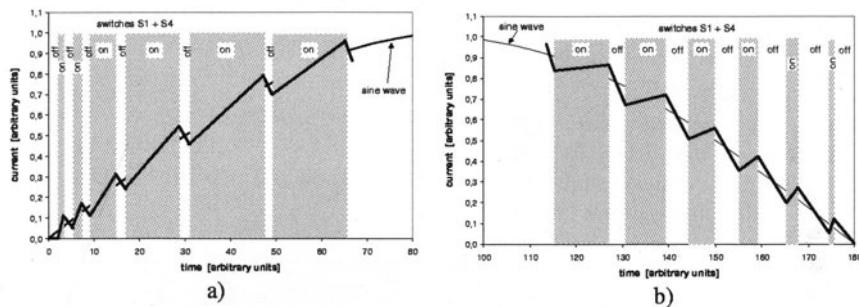


Fig. 15.5. Generation of a sine wave by using switches: the increasing part of the positive sine half-wave (a) and the decreasing part of the positive sine half-wave (b)

- a blocking-voltage capability in the off-state (blocking state) appropriate to the application (only a small leakage current is tolerable);
- low losses during the on-state (conducting state), which means a good conductivity during the current flow in order to minimize the requirements on the cooling system;
- low losses during the transition from the on-state into the off-state and vice versa.

Additional favorable properties are low demands on the gate circuitry and the possibility of switching frequencies up to 20 kHz, taking account of the losses which occur during the switching phases.

The necessary blocking-voltage capability of a power semiconductor device is determined by the voltage range typical of the respective application (Table 15.1).

Table 15.1. Applications and requirements on power semiconductor devices

Application	Blocking-voltage capability	Power semiconductor device
Automotive applications, servo drive systems	< 100 V	MOS transistor
Home appliances (e.g. air conditioning, washing machines, induction heating, microwave ovens); switch mode power supplies; power management (uninterruptible power supplies, welding systems)	≤ 600 V	MOS transistor Bipolar transistor Cool-MOS transistor IGBT
Industrial applications, e.g. motor drives for machine tools, robotics	≥ 1200 to 1800 V depending on the line voltage	IGBT
Special applications for high power (large motor drives for steel mills, electrical vehicles and wind power); High-voltage direct-current (HVDC) transmission	> 2000 V to 8000 V	Thyristor GTO (GTC) IGBT

15.2 Diode

As mentioned in the introduction, diodes (rectifiers) were the first power semiconductor devices realized.

The structure of a diode with one pn junction is shown schematically in Fig. 15.6. Between a high-doped p-region (p^+) and a high-doped n-region (n^+), a low-doped middle n-region (n or n^- , for low-doped or very low-doped, respectively) is inserted in order to sustain a voltage. If the polarity is chosen as indicated in Fig. 15.6a, the holes and electrons will be removed from both sides of the pn junction, leaving uncompensated doping atoms in the p-region on the left-hand side and in the low-doped middle region, and a space charge zone will build up. The diode is in a blocking state.

If the polarity is changed as in Fig. 15.6b, holes and electrons will be driven from both high-doped regions into the low-doped middle region, enhancing the conductivity dramatically: the conducting state now exists.

15.2.1 Blocking-Voltage Capability (Reverse or Blocking State)

In the blocking state, an electric field develops between the uncompensated (positive) donors n_{D+} in the n-doped middle region and the uncompensated (negative) acceptors n_{A-} in the p-doped border region (Fig. 15.7). The highest electric field exists at the pn junction, since this is the only point that

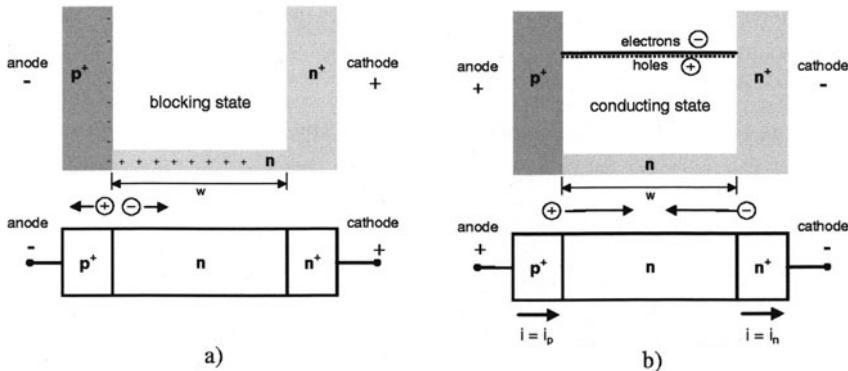


Fig. 15.6. Diode structure and doping levels: blocking state (a) and conducting state (b)

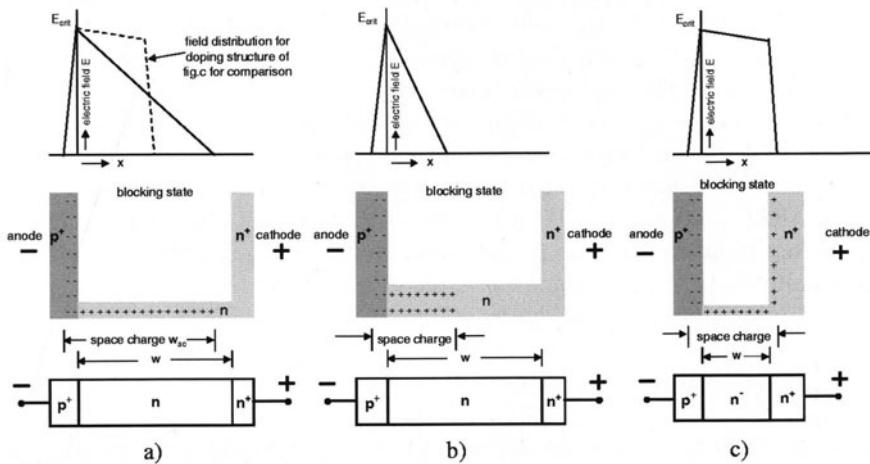


Fig. 15.7. Diode structure, doping levels and distribution of the electric field: (a) medium doping level of the middle region; (b) enhanced doping level of the middle region; (c) very low doping level of the middle region

all field lines between the positive and the negative charges have to pass, through.

The field distribution in a semiconductor device is described by the Poisson equation

$$\frac{d^2\varphi}{dx^2} = -\frac{\rho_{sc}}{\varepsilon \cdot \varepsilon_0}. \quad (15.1)$$

When this is taken together with the relation for the electric field E ,

$$E = \frac{d\varphi}{dx}, \quad (15.2)$$

and that for the space charge density ρ_{sc} , determined by the uncompensated doping atoms n_{doping} ,

$$\rho_{sc} = q \cdot n_{doping}, \quad (15.3)$$

where q is the electron charge, the gradient of the electric field is given by

$$\frac{dE}{dx} = -\frac{q \cdot n_{doping}}{\varepsilon \cdot \varepsilon_0}. \quad (15.4)$$

With $\varepsilon = 11.2$ for silicon and $\varepsilon_0 = 8.85 \times 10^{-14} \text{ A s}/(\text{V cm})$, the value $\varepsilon \varepsilon_0 \approx 10^{12} \text{ A s}/(\text{V cm})$ results.

From the relation between the voltage V and the electric field E ,

$$V = \int E \cdot dx, \quad (15.5)$$

the voltage across the p^+nn^+ structure can be calculated; it is expressed also by the area enclosed by the field distribution.

According to (15.4), the doping concentration n_{doping} determines the field gradient, whereas the sign of the space charge (negative for uncompensated negative acceptors n_{A^-} or positive for uncompensated positive donors n_{D^+}) is responsible for an increasing or decreasing field, respectively.

Figure 15.7a shows the space charge zone w_{sc} with the positive charge on the right-hand side of the pn junction and the negative countercharge on the left-hand side of the pn junction. In this case, according to (15.5), the maximum blocking voltage $V_{R\max}$ is given by the triangular area, with $E_{max} = E_{crit}$ and the space charge extension being w_{sc} :

$$V_{R\max} = E_{crit} \cdot \frac{w_{sc}}{2}. \quad (15.6)$$

As shown in Fig. 15.7b, if the doping in the middle region is enhanced by a factor of 2, only half the voltage can be built up across the device.

The most important contribution to the blocking-voltage capability comes from the low-doped middle region, because usually the doping of the adjoining border regions is a few orders of magnitude higher.

With decreasing doping level in the middle region, the field distribution changes from a triangular to a trapezoidal form (Fig. 15.7c), and the maximum voltage capability can be expressed – neglecting the very thin space charge zones in the high-doped border regions – as at least

$$V_{R\max} = E_{crit} \cdot w. \quad (15.7)$$

Now the whole middle region w is depleted, and the field on the right-hand side of the pn junction decreases in the high-doped n^+ -region. For the same blocking-voltage capability as in Fig. 15.7a, a reduced width of the middle region w can be used, which is favorable for the conducting state.

In silicon the maximum electric field E_{crit} is limited by the Zener effect for voltages $V_{R\max} < 10$ V and by an avalanche process for $V_{R\max} > 10$ V to values of about 300 kV/cm and 200 kV/cm, respectively. At these high electric fields additional carriers will be generated, originating from the silicon lattice, and a strong increase of the leakage current will occur. Either these carriers are extracted directly from the lattice (Zener effect) or electrons are accelerated to such an extent that fixed electrons in the lattice are knocked off their sites and are then accelerated, resulting in a chain reaction (avalanche effect).

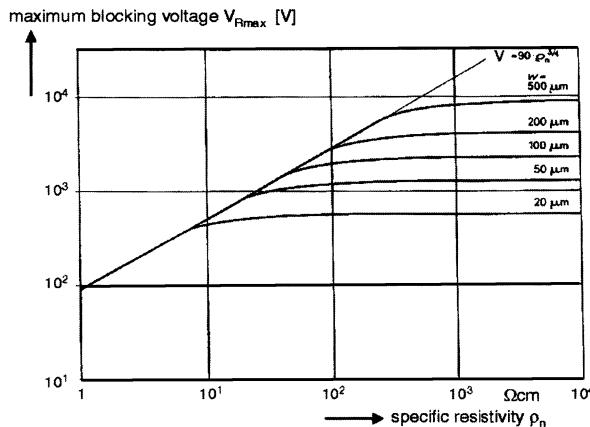


Fig. 15.8. Maximum blocking-voltage capability of a p^+nn^+ (diode) structure

Figure 15.8 shows the maximum blocking voltage $V_{R\max}$ for p^+nn^+ structures depending upon the n-doping and the width of the middle region. The saturation of the blocking capability for a given thickness w of the middle region in accordance with (15.7) can be observed. On the abscissa, the resistivity ρ_n of the n-doped middle region is used,

$$\rho_n = \frac{1}{\sigma} = \frac{1}{q \cdot \mu_n \cdot n_{D^+}}. \quad (15.8)$$

Here, the conductivity σ is given by

$$\sigma = q \cdot \mu_n \cdot n_{D^+}, \quad (15.9)$$

where n_{D^+} is the doping concentration of the low-doped middle region, μ_n is the mobility of the electrons and q is the electron charge.

It can be noticed that the increase of the blocking-voltage capability is not directly proportional to the resistivity as would be expected from Fig. 15.7. With a growing space charge zone, a high electric field exists for a longer distance and the avalanche effect occurs at slightly decreasing values of E_{crit} .

15.2.2 Conducting (Forward) State

In the diode's conducting state, the voltage drop V_F across the diode during current flow should be as low as possible in order to reduce the losses which determine the heating up of the device and therefore define the cooling system necessary.

In contrast to the blocking state, with its lack of holes and electrons resulting in a space charge zone, holes and electrons are now swept into the middle region and a new balance is possible with a neutrality of charge (Fig. 15.6b),

$$n = p + n_{D+}, \quad (15.10)$$

the negative charge of the electrons (n = concentration of electrons) approximately compensating the positive charge of the holes (p = concentration of holes). The doping concentration n_{D+} in the middle region can be neglected because it is 10 to 100 times lower than the concentrations n and p .

In the n^+ -region on the right-hand side in Fig. 15.6b, the current is an electron current i_n , but in the p^+ -region on the left-hand side the current has to be a hole current i_p . In the middle region, the whole current has to be transformed from an electron current into a hole current by the recombination of electrons and holes. Since the recombination rate is proportional to the carrier concentrations, the concentrations are also dependent on the value of the current which has to be transformed. In addition the recombination rate is inversely dependent on the carrier lifetime in this region.

If the recombination surplus is denoted by R – the difference between the recombination rate and the generation rate – the following equation for the current density i is valid:

$$i = q \cdot \int R \cdot dx = q \cdot \frac{n}{\tau_{hl}} \cdot w, \quad (15.11)$$

where the concentration of the carriers is $n = p$, the width of the middle region is w , and τ_{hl} is the high-level lifetime of the carriers in the middle region of the diode.

The generation rate is responsible for the leakage current in the blocking state, which is below 10^{-2} A/cm^2 and can be neglected in the forward (conducting) state, where forward current densities are about 100 A/cm^2 .

For a constant current density i and therefore a constant recombination surplus R , the concentrations of the electrons n and holes p can be kept lower when the high-level lifetime τ_{hl} has a low value, according to (15.11). But for lower concentrations the conductivity becomes worse, according to the following equation:

$$\sigma = q \cdot \mu_n \cdot n + q \cdot \mu_p \cdot p. \quad (15.12)$$

Figure 15.9 shows the simulated carrier concentrations $n = p$ in the middle region of a diode for different carrier lifetimes τ_{hl} . If the lifetime τ_{hl} is reduced by a factor of 4, the carrier concentration (and the stored charge) is also

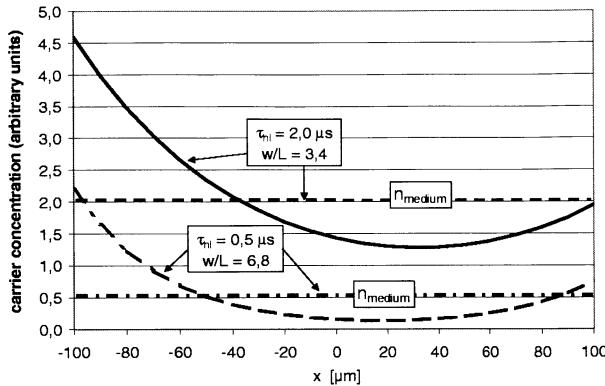


Fig. 15.9. Calculated carrier distributions in the middle region of a diode in the conducting state for different high-level lifetimes τ_{hl}

reduced by a factor of 4 (indicated by the medium concentrations n_{medium}). But the lowest concentration value in the middle region is reduced by a factor of 10 because of a stronger curvature of the carrier distribution in the case of a lower τ_{hl} . Since the lowest concentration is responsible for the highest contribution to the forward voltage drop V_F , this effect increases the losses in the conducting state more than proportionally.

For bipolar devices, the ratio

$$\frac{\text{width of the middle region}}{\text{diffusion length of the carriers}} \frac{w}{L}$$

is an important measure which determines the curvature of the carrier distribution.

The diffusion length L is given by the equation

$$L = \sqrt{D \cdot \tau_{hl}} = \sqrt{\frac{\mu \cdot k \cdot T}{q} \cdot \tau_{hl}}, \quad (15.13)$$

where the mobility of the carriers is μ , the temperature in degrees Kelvin is T , the electron charge is q and the Boltzmann constant is k ($k = 1.38 \times 10^{-23} \text{ Ws/K}$). The mobility μ for the high-level condition $n = p$ is a combination of the mobilities of holes μ_p and electrons μ_n ,

$$\mu = 2 \frac{\mu_n \cdot \mu_p}{\mu_n + \mu_p}, \quad (15.14)$$

and is derived by solving the continuity equations.

Figure 15.10 depicts the dependence of the forward voltage drop V_F in the conducting state on the width at a fixed lifetime τ_{hl} (a) and on the lifetime τ_{hl} at a fixed width w (b) for a typical diode structure. If the recombination

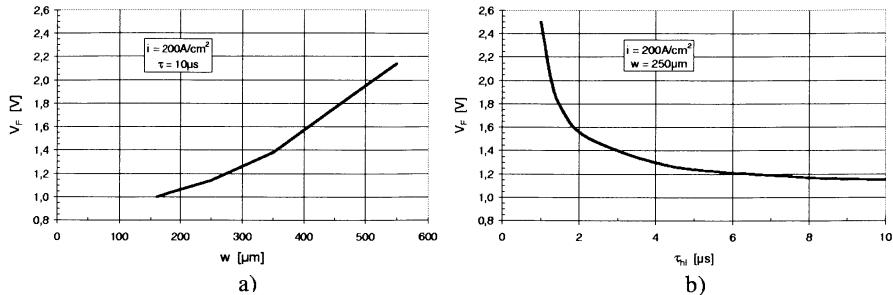


Fig. 15.10. Forward voltage drop V_F of a diode: dependence on the width of the middle region w (a) and dependence on the high-level lifetime τ_{hl} (b)

in the border regions can be neglected, the amount of stored carriers in a swamped middle region w of a diode with an area A is equal to

$$Q_S = q \cdot n \cdot w \cdot A \quad (15.15)$$

and can be measured by extracting the stored charge Q_S .

From (15.11), introducing the forward current I_F ,

$$I_F = i_F \cdot A = q \cdot \frac{n}{\tau_{hl}} \cdot w \cdot A, \quad (15.16)$$

together with (15.15), a measure of the high-level lifetime τ_{hl} can be obtained,

$$\tau_{hl} = \frac{Q_S}{I_F}, \quad (15.17)$$

if the measurement is carried out in a time interval short in comparison with the high-level lifetime τ_{hl} .

In fast switching applications, the stored charge has to be kept low to avoid unacceptably high reverse recovery currents (see below) and to reduce the losses during the switching process.

By introducing recombination centers with energy levels located in the middle of the band gap, the high-level lifetime τ_{hl} can be controlled. A lifetime doping can be done by a diffusion process with gold or platinum or by irradiation with electrons, protons or helium atoms. The lifetime doping also influences the generation process, which dominates the leakage current in the blocking state. Especially at higher temperature, this leakage current of the device can be enhanced in comparison with devices without a lifetime doping.

Typical forward voltage drops V_F of power silicon diodes are in the range 1 V–4 V at 50–200 A/cm², depending on the blocking-voltage capability and the application. A voltage–current characteristic for a 1700 V/300 A diode used in fast switching circuits today is shown in Fig. 15.11. With increasing temperature, the high-level lifetime increases and higher concentrations of

holes and electrons are necessary, according to (15.11). A better conductivity with a lower forward voltage drop V_F is obtained, but the result is also a higher stored charge. On the other hand, the mobility of the carriers decreases with increasing temperature and the forward voltage drop grows. Depending on the rate of increase of the carrier concentrations and the rate of decrease of the mobility of the carriers, either a negative or a positive temperature coefficient results for the forward voltage drop V_F at a given current. A good balancing of these two effects can be recognized in Fig. 15.11.

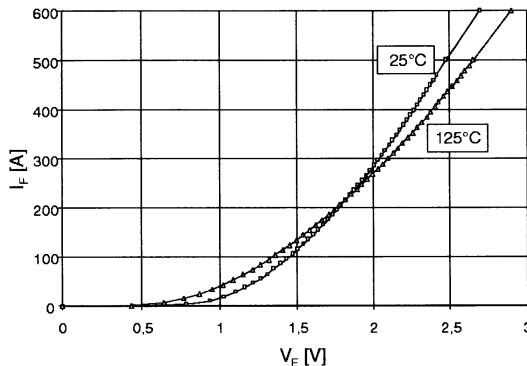


Fig. 15.11. Voltage V_F -current I_F characteristics of a 1700 V/300 A fast switching diode in the conducting state at 25°C and 125°C

15.2.3 Dynamic Behavior

Since power diodes are used as switching devices, the dynamic situation during switching from an on-state into an off-state and vice-versa has to be considered in addition to the static behavior.

Turn-Off

If the polarity of the voltage across the diode is changed from a forward bias to a reverse bias, the carriers “swamped” into the middle region during the forward (conducting) state have to be removed in order to establish the reverse blocking voltage.

In Fig. 15.12, the sweeping-out process (also called the reverse recovery process) is shown schematically. In the case of fast switching – a fast change of the polarity – the current in the reverse direction is limited at the beginning only by the external circuit parameters (resistance or inductance) and not by the diode itself. In the still-swamped middle region, where $n = p$, the current is dominantly a field current, the sum of the electron current i_n and the hole current i_p given by

$$i = i_n + i_p = q \cdot n \cdot \mu_n \cdot E + q \cdot p \cdot \mu_p \cdot E. \quad (15.18)$$

Since $\mu_n \approx 3\mu_p$, the field current of the electrons i_n is about three times higher than the hole current i_p :

$$i_n \approx 3 \cdot i_p. \quad (15.19)$$

The electrons flow towards the positive potential in the high-doped n^+ -region, and the holes are driven towards the negative potential in the high-doped p^+ -region. Because electrons cannot be supplied from the p^+ -region, a zone develops in the middle region near the pn junction where no electrons are present. In this region, the whole current has to be carried by the holes, which, together with uncompensated donors, establish a positive space charge zone. The negative countercharge (acceptors n_{A-}) is located in the high-doped p^+ -region. On the right-hand side in Fig. 15.12, a hole-free zone exists and the electrons can build up a negative space charge or they can compensate the positive donors, with an ohmic zone as a result. Because of the different mobilities of electrons and holes, more electrons than holes are removed in a given time interval (indicated by the flow of three electrons and one hole in the still-swamped region). The development of the positive space charge zone on the left-hand side in Fig. 15.12 starts earlier, and this zone grows faster than the zone on the right-hand side in this figure.

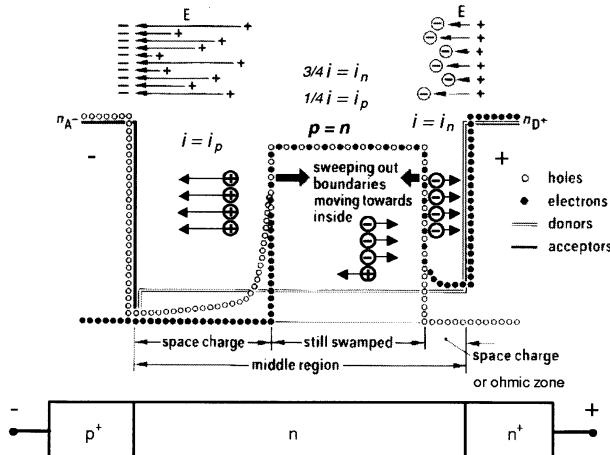


Fig. 15.12. Diode structure and flow of the carriers in the middle region of a diode during the reverse recovery process

During the flow of the reverse current these zones move towards the inside of the middle region. The emergence of a space charge zone is paralleled by the building up of a blocking voltage across the diode. Since the switching

time is usually short in comparison with the high-level lifetime τ_{hl} , the recombination of carriers can often be neglected during the switching process. But the lifetime τ_{hl} is decisive for the starting condition, and the maximum reverse recovery current – comparable in size to the forward current – is also strongly influenced.

A typical current waveform for a 1700 V/200 A fast switching diode is shown in Fig. 15.13 (labeled as “conventional diode”) for a forward current of 200 A commutated in about 0.4 μ s to the maximum value of the reverse recovery current of 220 A, with a subsequent soft return of the current to zero.

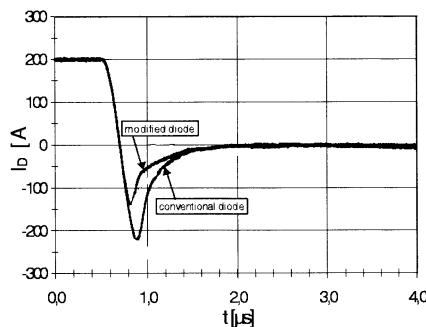


Fig. 15.13. Reverse recovery current waveforms of 1700 V/200 A diodes having different carrier distributions in the middle region

Together with the increasing blocking voltage across the device, not shown in this picture, peak losses in the range of a few times 100 kW/cm^2 are possible and have to be sustained by the device.

The waveform of the reverse recovery current is very important as well. During the reverse recovery process, a sharp decrease of the reverse recovery current – a snap-off behavior – has to be avoided, since dangerous voltage spikes might occur. To avoid a snap-off, the diode has to be designed in such a way that there is a certain amount of charge present in it at all times until the build up of the blocking voltage is finished. If a high reverse recovery current appears, the remaining charge will be removed very fast by the flowing reverse recovery current and a sudden current decrease may occur. A detailed description of the reverse recovery process in diodes can be found in [14].

Turn-On

Usually the turn-on of a diode is not a critical process, but it should be mentioned that the swamping with carriers needs some time. At the first moment of the process the conductivity of the diode is determined only by the doping concentration of the middle region, resulting in a resistance of about $10\text{--}100 \Omega$. During a fast current increase, the voltage across the device may

reach 10 V to a few hundred volts, which is dangerous to a device arranged in parallel with the diode that does not have a blocking capability in the relevant direction.

15.2.4 Trends

In order to optimize the voltage drop and the dynamic behavior of a diode, development during recent years has been focused on the manipulation of the carrier concentration distribution in the middle region. To obtain an acceptable maximum reverse recovery current under fast switching conditions, it is most important to reduce the carrier concentration $n = p$ in the middle region near to the p^+ -region. This can be achieved either by enhancing the recombination in the high-doped p -region, e.g. by irradiation with protons or helium, or by reducing the emitter efficiency by lowering the doping of this p^+ -region. The results are shown in Figs. 15.13 and 15.14. In Fig. 15.13, the reverse recovery behaviors (current waveforms) of two diodes which have different carrier concentration distributions in the middle region are compared. With a special laser technique [15], the carrier concentration distributions for these two diodes were measured, and the difference is recognizable in Fig. 15.14.

To improve the behavior of diodes an additional gate terminal has been proposed. But more complex control circuits are necessary for such devices. Distinct improvements will be possible if other materials are used, e.g. silicon carbide, which today is used only for special applications in the low-voltage, low-current range (e.g. for switch mode power supplies with switching frequencies in the range of 100 kHz). At present, the most important disadvantages of this material are the very high price and defects in the material, which limit the chip size to only a few millimeters.

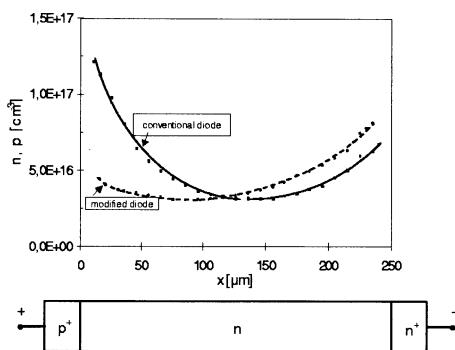


Fig. 15.14. Carrier distributions in the middle region, and structure of the diodes shown in Fig. 15.13

15.3 Thyristor

15.3.1 Behavior in Principle

As shown schematically in Fig. 15.15 the thyristor is a three-terminal device with terminals referred to as the anode, cathode and gate. Four layers doped alternately n and p result in three pn junctions (J1, J2, J3).

If a polarity is applied as indicated in Fig. 15.15a, the junctions J1 and J3 will be in a blocking state; junction J2 is biased in the forward direction but the blocking junctions J1 and J3 dominate: this is the reverse blocking state. Junction J3 determines the blocking-voltage capability because at junction J1, the adjoining regions are usually high-doped and only a low blocking voltage can be built up (10 V–50 V).

When the polarity is changed as indicated in Fig. 15.15b, the pn junction J2 is then in a blocking state and the pn junctions J1 and J3 are biased in the forward direction: this is called the forward blocking state. A comparison of Figs. 15.15a and 15.15b shows that in principle the same blocking capabilities are possible, because in both cases the same low-doped middle region is responsible for the blocking voltage.

In the forward state, an additional condition is possible when carriers coming from the outer n- and p-type regions (where junctions J1 and J3 are forward biased) swamp the blocking pn junction J2 (Fig. 15.15c). Usually this can be initiated by a third terminal, the gate. When a positive potential is applied to the gate terminal with respect to the cathode, the pn junction J1 is biased more strongly in the forward direction and electrons flow towards the anode (Fig. 15.15c). At the anode junction J3, a hole current is triggered by the electrons, which flows towards the cathode. This hole current in turn is responsible for a higher electron current from the cathode. If the emitter

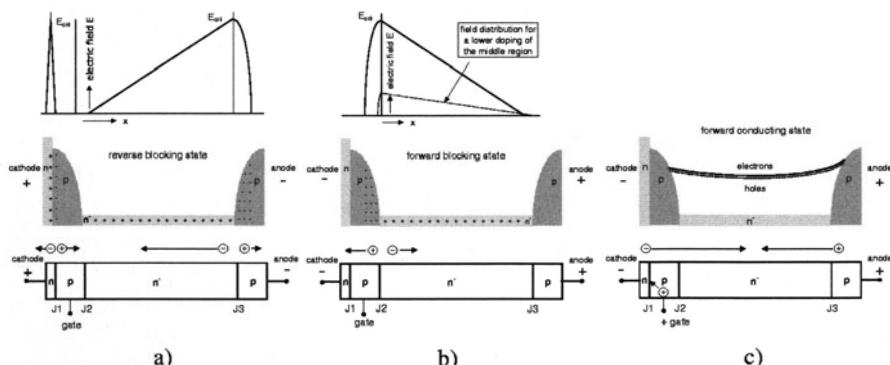


Fig. 15.15. Structure, doping levels and distribution of the electric field of a thyristor in the reverse blocking state (a), in the forward blocking state (b) and in the forward conducting state (c)

efficiencies at the junctions J1 and J3 are high enough, a regenerative process can be maintained, resulting in a fully swamped middle region.

A more realistic cross section of a thyristor is shown in Fig. 15.16. Usually the gate terminal is arranged to be in the middle of the structure and the emitter regions extend to the periphery.

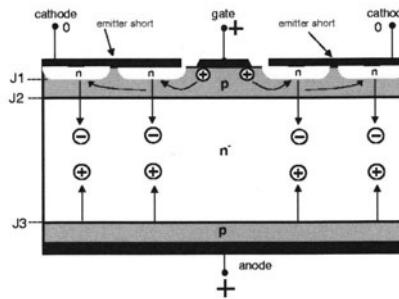


Fig. 15.16. Cross section of a thyristor structure with the current flow in the conducting state

15.3.2 Blocking-Voltage Capability

For the blocking-voltage capability of an n^-p structure in the forward and the reverse direction, the critical field strength E_{crit} and the width of the low-doped n^- middle region are important again (Fig. 15.15). But as the electrical field approaches the opposite forward-biased pn junction, more and more holes are collected and a large increase of the reverse current limits the blocking-voltage capability. Especially if the middle region is very low-doped, the gradient of the electric field will be small and a punch-through of the field to the opposite p-layer will be possible at a low maximum electric field long before the critical field strength E_{crit} is reached. The result is a low blocking-voltage capability (see, for example, the dashed lines for the electric field in Fig. 15.15b in the forward blocking state).

Usually the p-regions of a thyristor are produced by diffusion processes, resulting in pn junctions J1 and J3 at a depth of 30–100 μm below the silicon surfaces, with decreasing doping concentrations from the surface towards the inside, as indicated in Fig. 15.15. But again the low-doped middle region is the factor that determines most substantially the blocking-voltage capability.

The highest value of the blocking voltage of an (n)pnp structure is obtained at a well-defined doping concentration (or a well-defined resistivity of the silicon material) and depends also on the width of the middle region, as shown in Fig. 15.17 [16].

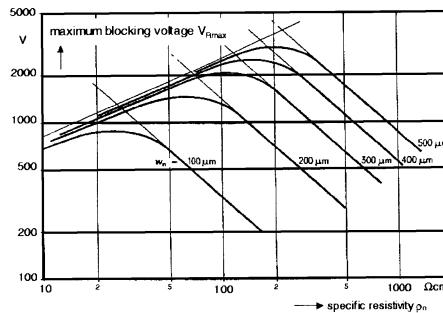


Fig. 15.17. Maximum blocking-voltage capability of a thyristor (npnp structure)

In contrast to the blocking-voltage capability of a diode shown in Fig. 15.8, a decreasing blocking-voltage capability may occur if the resistivity is increased without adapting the width of the middle region.

In accordance with Fig. 15.17, the spread of the doping of the low-doped n region determines also the spread of the blocking voltage. In order to obtain a given blocking-voltage capability with a good yield in the production of thyristors, structures with a large spread of the doping need a larger width of the low-doped n region than structures which have a small spread of the doping of the silicon material. With the introduction of neutron transmutation doped silicon, the doping variations of $\pm 20\%$ typical for silicon material produced by the conventional zone refining process could be reduced to $\pm 5\%$ [17]. The improvement of the characteristics of thyristors obtained by using a reduced spread of the doping of the starting silicon material and having also a reduced width is shown in [18].

A further reduction of the width of the middle region is possible if an n-doped buffer layer is inserted between the n⁻-region and the p anode layer. If a lower-doped n⁻-region is used the electric field will be stopped in the n buffer layer similarly to the situation in Fig. 15.7c for the diode, but the reverse blocking-voltage capability has to be sacrificed. But it is necessary to ensure that the emitter efficiency is not reduced too much by the higher-doped buffer layer.

15.3.3 Conducting State

The low-doped middle region of a thyristor is swamped with holes and electrons in a similar way as was discussed for the diode; this is indicated in Fig. 15.15c. If the thickness of the low-doped region and the area are made equal for a diode and a thyristor, about the same voltage drops result. A small difference appears only in the low-current range, caused by the additional p-region between the n cathode region and the low-doped middle region [19].

For a thyristor with a symmetrical blocking characteristic, the width of the middle region of a thyristor is usually larger for the same blocking-voltage

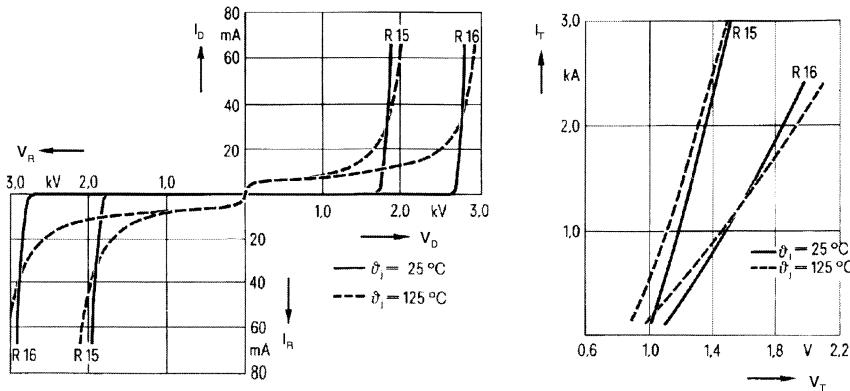


Fig. 15.18. Blocking-voltage characteristics (left) and forward voltage characteristics (right) of thyristors with 1650 V and 2500 V blocking-voltage capability for a current rating of 1400 A and 800 A, respectively, at 25°C and 125°C

capability compared with a diode because punch-through has to be avoided. Therefore the forward voltage drop V_T of the thyristor in the conducting (triggered) state is higher than the V_F for a diode.

With increasing blocking-voltage capability, accompanied by an adapted increasing width of the middle region, the forward voltage drop also increases.

In Fig. 15.18, the blocking and conducting characteristics of thyristors (in this example, devices made from silicon 55 mm in diameter) with blocking-voltage capabilities of 1650 V (R15) and 2500 V (R16) are compared at 25°C and 125°C, showing a higher forward voltage drop for the thyristor R16 with the higher blocking-voltage capability.

15.3.4 Dynamic Behavior

Turn-Off

For the dynamic behavior of the thyristor, the turn-off time t_f is an important figure. It is defined as the time interval which has to elapse after the current has changed from a forward (conduction) state (Fig. 15.15c) to the reverse blocking state (Fig. 15.15a) and until a voltage in the forward direction can be applied again. In a manner comparable to the behavior of a diode, a reverse recovery current appears in a thyristor structure as well after the reversal of the supply voltage.

The holes flow to the anode and the electrons are driven towards the cathode but the electrons now have to pass through the pn junction J2 and maintain a delivery of holes into the middle region (Fig. 15.19). As in the diode, a space charge zone develops at junction J3, but in contrast to the diode a full removal of the carriers is hindered by the continuing delivery of holes, which is proportional to the corresponding current flow. After the voltage is established across the thyristor in the reverse direction, the current decreases

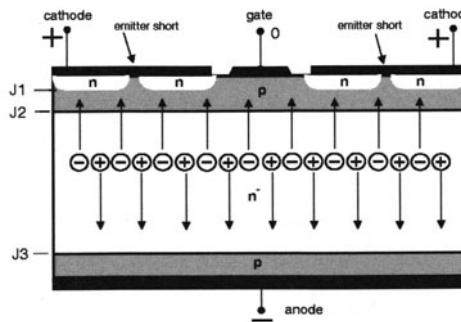


Fig. 15.19. The flow of electrons and holes after a change of the polarity from the forward to the reverse direction

and, basically, the lifetime τ_{hl} is responsible for the further removal of the carriers and dominates the turn-off time t_f , which is about 7–10 times the high-level lifetime τ_{hl} . This lifetime has to be well adapted to the requirements of the application because a reduction of the carrier lifetime increases the forward voltage drop V_T , similar to what was shown for diodes in Fig. 15.10b.

Turn-On

A gate current between the gate terminal and the cathode initiates the turn-on of the thyristor (Fig. 15.16). A gate current pulse of the order of 1 A with a duration of about 10 μ s is usually sufficient to establish the regenerative process.

At the beginning, only the cathode regions adjoining the gate electrode are turned on, and relatively small areas are conducting the current (Fig. 15.20). Since the current is determined by the circuit parameters, high specific power losses exist and the possible destruction of the device is imminent if a high current increase di/dt is applied (Fig. 15.21). The turned-on plasma region spreads with a velocity of 20–200 $\mu\text{m}/\mu\text{s}$ over all the cathode area [20, 21].

For applications with high di/dt values, special cathode and gate constructions have been developed with less extended emitter regions and which take advantage of the long rims of interdigitated structures. Because such

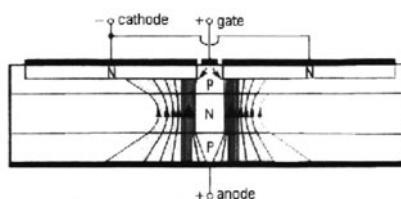


Fig. 15.20. Inhomogeneous current distribution at the beginning of the turn-on of a thyristor

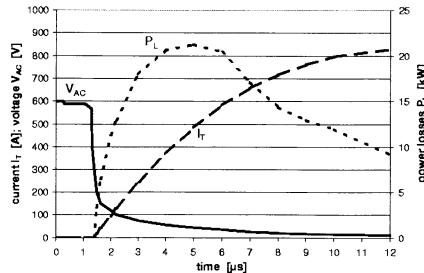


Fig. 15.21. Waveforms of the voltage V_{AC} the current I_T and the losses P_L during the turn-on of a thyristor

enlarged cathode rims need significantly higher gate currents for a uniform turn-on of the whole rim, a so-called amplifying gate [22] is used, which turns on first and then the flowing load current is effective as a gate current that turns on the enlarged cathode rim.

In Fig. 15.22, the spreading of the turned-on area can be observed by detecting the infrared radiation resulting from the recombination of electrons and holes during the turn-on process. By special preparation of the thyristor and by using a photomultiplier, the infrared radiation can be made visible [23]. A turn-on from a 500 V blocking voltage with a current increase $di/dt \approx 100 \text{ A}/\mu\text{s}$ up to 600 A is shown for a 900 V thyristor with a turn-off time of 10 μs , designed with an amplifying gate and an interdigitated finger structure. The amplifying-gate thyristor turns on first (top center), making possible a uniform turn-on of the main thyristor with large cathode rims in the following microseconds. After 25 μs the growing together of the bright areas can be recognized for two fingers (prepared with observation holes in the metallization) in the picture at the bottom right.

In order to handle high dV/dt values and the displacement current that appears in the forward blocking direction without initiating the regenerative process, a large part of the displacement current (holes) has to be kept away from the n-emitter. For this reason, so-called emitter shorts are arranged in the emitter areas (Fig. 15.16), and the hole current has a chance to flow directly to the cathode electrode, bypassing the n-emitter. The size and arrangement of these emitter shorts have to be well adjusted, since the gate trigger current and the spreading of the turned-on area may also be influenced in an unfavorable way. A survey of specific problems in bipolar power devices is given in [24].

15.3.5 Trends

Today, thyristors are being produced with blocking capabilities up to $> 8 \text{ kV}$. In the lower-voltage range the thyristor is being replaced more and more by

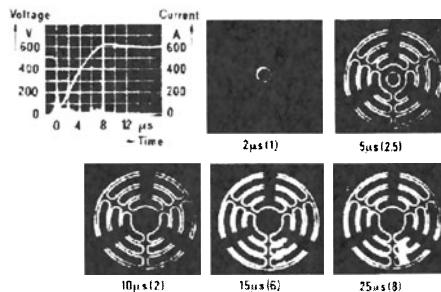


Fig. 15.22. Waveforms of current and voltage (top left), and infrared recombination radiation during the turn-on of a thyristor with an interdigitated gate structure at various times

the IGBT, but in high-voltage, high-power applications (e.g. HVDC, high-voltage direct-current transmission) it is still the most important device.

New developments permit triggering by light, which is most interesting for applications in the highest voltage range because isolation problems can be solved more easily [25].

15.4 GTO (Gate Turn-Off Thyristor)

15.4.1 Behavior in Principle

The GTO is a product of the continuous development of the thyristor structure aimed at making possible a turn-off by a gate current at any time as well as a turn-on. Concerning the blocking capability, the same physical relations hold as for thyristors, including the possibilities of symmetrical and asymmetrical structures.

The doping concentrations of the p- and n-regions are different from those of a thyristor in order to optimize the turn-off behavior. The geometrical designs of the cathode and gate regions are different as well, as indicated in Fig. 15.23. Whereas the width of the n-cathode region for a thyristor has dimensions in the range of 2–20 mm including finger structures, the n-cathode region of the GTO is divided into fingers, each having a width of only a few hundred microns.

In the conducting state, the carrier concentrations of electrons and holes swamped into the middle region, which are again influenced by the carrier lifetime and by the width of the middle region, are responsible for the forward voltage drop.

15.4.2 Dynamic Behavior

The turn-on process is initiated by a positive potential at the gate with respect to the cathode, resulting in a forward current at the pn junction J1

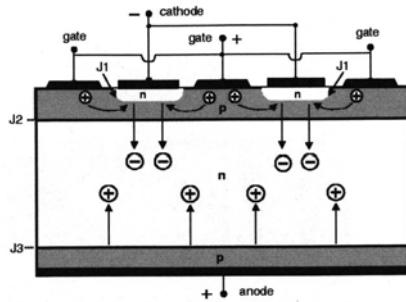


Fig. 15.23. Cross section of a GTO (gate turn-off thyristor) showing the flow of holes and the electrons in the conducting state

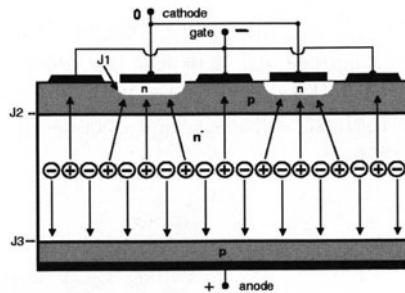


Fig. 15.24. The flow of electrons and holes during the turn-off of a GTO

(Fig. 15.23). Electrons flow from the cathode (n-emitter) towards the anode, starting the regenerative process, with a swamped middle region as a result, similar to that indicated in Fig. 15.15c for the thyristor.

To initiate the turn-off process, part of the load current has to be commutated into the gate circuit by applying a negative potential at the gate with respect to the cathode without changing the polarity at the anode and cathode (Fig. 15.24). Owing to the small dimensions of the emitter regions, holes are more or less kept away from the n-emitter and the regenerative process can be hindered in such a way that a turn-off behavior is obtained. In order to interrupt the regenerative process, quite a high share of the load current has to be transferred into the gate circuit (10–40%). But in the gate circuit only low (blocking) voltages (about 20–30 V) are available for the current transfer, owing to the high-doped regions which make up the pn junction J1. In the case of high-current devices, high currents also have to be handled in the gate circuit.

During the turn-off process the GTO tends to build up areas of current filaments with a high power density, and therefore the risk of destruction is high. So usually the increase of the voltage across the device has to be limited

by a snubber circuit comprising a capacitor, a resistor and a diode arranged in parallel with the device, as is also done with thyristors.

15.4.3 Trends

At the beginning of the 1990s an improved GTO – the GCT (gate-commutated thyristor) – was introduced [26]. By optimizing the doping concentrations and the geometry on the cathode side, it is possible to commutate the whole load current into the gate circuit within 1 μ s, which is necessary to avoid the formation of filaments. In order to realize such a switching behavior, a low-inductance gate circuit has to be developed, but a snubber circuit, which is necessary for the GTO, can be omitted.

Today, the GTO is used almost exclusively for applications with high-voltage, high-current requirements (e.g. traction and HVDC). Special developments such as the reverse-conducting GTO, which includes an integrated diode, have not found widespread use.

15.5 Bipolar Transistor

Like the GTO, the bipolar transistor takes advantage of a capability to turn on and turn off at any time. Although many papers about bipolar transistors were published in the USA, the use of bipolar transistors in power applications was pioneered in Japan by the introduction of modules with bipolar transistor chips arranged in parallel on an isolating substrate in order to meet the requirement for high-current devices. Such devices were used to a great extent in the 1970s to regulate air conditioning systems, which were more important in the climate of Japan.

15.5.1 Behavior in Principle

The bipolar transistor has a three-layer structure, e.g. an npn structure with an emitter, a base and a collector terminal. For power bipolar transistors, a

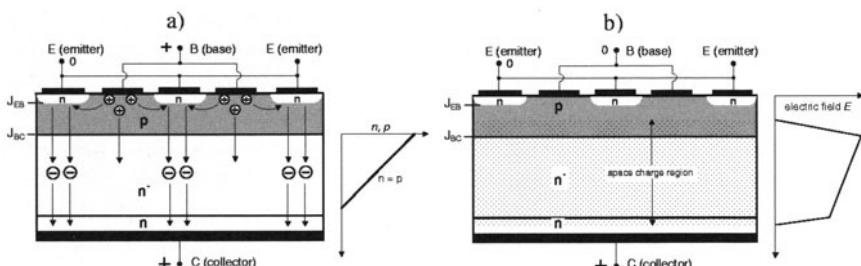


Fig. 15.25. Power bipolar transistor in the conducting state (a) and in the blocking state (b)

structure of the kind shown in Fig. 15.25 has been established, with a thin p-base region and with a collector having a low-doped n⁻-region in front of a higher-doped n-region. The adjoining regions at junction J_{BC} determine the blocking capability. As discussed earlier, at least one layer has to have a low doping and an extension large enough to take up the space charge region.

In the conducting state, the junction J_{EB} is biased in the forward direction by a positive potential at the base terminal with respect to the emitter (Fig. 15.25a). The base current (holes) between the base and emitter triggers a flow of electrons from the emitter to the collector. In principle, the geometry on the emitter side has a finger structure, similar as discussed in the case of the GTO.

As shown in Fig. 15.25b, in the blocking state the space charge zone can spread into the low-doped n⁻ collector region in order to sustain the blocking voltage.

15.5.2 Conducting State

The power bipolar transistor is usually used in an emitter configuration (Fig. 15.25a) in which a small current in the base circuit can control a large current in the load circuit between the collector and the emitter.

The current gain is expressed by

$$\alpha_E = h_{FE} = \frac{\text{load current at the collector}}{\text{base current}} \frac{I_C}{I_B}, \quad (15.20)$$

describing the base current necessary for a certain load (collector) current.

Typically, for power bipolar transistors, at large load currents the low-doped n⁻ collector region is more or less swamped by carriers, but the load current causes a voltage drop in the remaining low-doped collector region. A forward biasing of the junction J_{BC} results and a flow of holes into the low-doped n-region is possible. A concentration distribution $n = p$ with a gradient from the emitter towards the collector appears, as indicated in Fig. 15.25a on the right-hand side. Whereas the electron diffusion current and the electron field current have the same direction and together form the load current, the diffusion and field currents of the holes exactly compensate each other. The holes delivered by the base current are needed only to support the neutrality, since continuous recombination takes place in the swamped region.

With increasing load (collector) current density, the concentration gradient must increase, resulting in higher concentration values. More and more base current is used to maintain neutrality in the swamped collector region, and therefore the share of the base current available for triggering the flow of electrons from the emitter is reduced and the current gain h_{FE} decreases. Consequently the base current needed for controlling the transistor increases, and for a transistor having a blocking-voltage capability of about 1000 V, 10–30% of the value of the load current has to be spent as a base current.

A reduction of the base current can be obtained by use of Darlington configurations, but disadvantages concerning the voltage drop in the conducting state and the switching behavior have to be accepted.

15.5.3 Blocking-Voltage Capability

Generally, the same relations as discussed for the thyristor are valid. But in principle, for pnp or npn structures, it is necessary to consider the case of an open-base condition (V_{CEO}), as well as the case where the emitter and base terminals are connected (V_{CBO}).

If an open base exists, holes of the leakage current generated in the space charge region around the collector–base junction J_{BC} have to pass through the forward-biased emitter–base junction J_{EB} , and the electrons emitted from this junction enhance the leakage current. Both types of carriers are accelerated in the space charge zone, resulting in an amplification of the current. The additional holes produced, flowing again to the emitter terminal, are in turn responsible for a higher electron current from the emitter.

The blocking voltage with an open base V_{CEO} is expressed by

$$V_{CEO} = \frac{V_{CBO}}{\sqrt[4]{h_{FE} + 1}}. \quad (15.21)$$

With common current gains in the low-current range of $h_{FE} = 10\text{--}20$, the blocking-voltage capability is reduced to about half the value V_{CBO} possible at the junction J_{BC} without an emitter region or with the emitter and the base regions connected. In this latter case the holes generated in the space charge region can bypass the emitter, facilitating a higher blocking-voltage capability.

The behavior just described exists in principle for all pnp and npn structures but is more pronounced for bipolar transistors because, here the base widths are very thin, being designed for a high current gain, and the emitter efficiency is not distinctly reduced, for example by emitter shorts as discussed earlier for the thyristor structure.

15.5.4 Dynamic Behavior

In order to guarantee a homogeneous forward bias of the emitter junction, the emitter regions are separated into fingers with a width smaller than 0.5 mm as schematically shown in Fig. 15.25. During the turn-on phase the low-doped collector region is swamped with carriers until the stationary distribution is established (Fig. 15.25a).

Severe problems may occur during the turn-off process, which is triggered by a negative potential at the gate terminal with respect to the emitter terminal. The base current is now flowing in the opposite direction compared with the situation during turn-on. The base current produces a voltage drop along

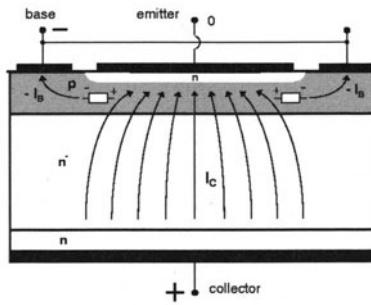


Fig. 15.26. The pinching of the current in a bipolar transistor during turn-off

the base layer resistance, leading to less forward biasing or even a reverse biasing of the rims of the emitter fingers, one of them magnified in Fig. 15.26. The current is more and more restricted to the inner regions of the emitter while the full load current is still flowing. Enhanced current densities influence the field distribution in an unfavorable way, and so-called second breakdown effects – described in numerous publications – may limit the switch-off capability of the bipolar transistor [27]. The problems concerning turn-off are more severe in Darlington configurations.

With negative base currents of 20–30% of the load current, the current crowding is less critical and also the switching losses are low, but perhaps longer storage times have to be accepted.

More detailed descriptions of the design of bipolar transistors and the behavior during turn-off can be found in [28] and [29, 30], respectively.

15.5.5 Trends

The severe problems concerning second breakdown effects could be solved to a large extent by introducing emitters with very small dimensions ($5\text{ }\mu\text{m}$ instead of $100\text{--}300\text{ }\mu\text{m}$ for the width of the fingers) [31]. But one remaining disadvantage of the bipolar transistor is its requirement for a continuous base current, which increases the cost of the base circuit. A second disadvantage is the necessity for a Darlington structure for blocking voltages higher than 600 V , with its enlarged voltage drop in the conducting state and with an unfavorable behavior during turn-off. Although the bipolar transistor was the first device which had a short-circuit capability, this behavior is limited to voltages below 1000 V .

Today, bipolar transistor modules are used mainly in home appliances and consumer electronics for applications in the low and medium power ranges at line voltages of 110 V and 220 V .

It can be stated in summary that the bipolar transistor opened the door to the introduction of modular construction and to the development of circuitry with turn-on and turn-off possibilities. However, in practice, the device has

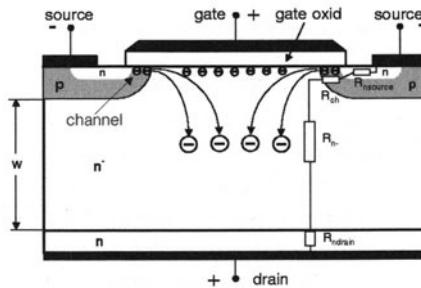


Fig. 15.27. The flow of electrons and holes in an MOS transistor in the conducting state

no physical potential to extend the blocking capability substantially higher than 1500 V, because of the additional requirements on high-voltage power semiconductor devices.

15.6 MOS Transistor (Metal–Oxide–Silicon Transistor)

15.6.1 Behavior in Principle

At the end of the 1970s, a structure was proposed to make use of the advanced technology used for integrated circuits [7,8]. Much-improved lithography with line widths down to the micron scale, the deposition of different layers such as polysilicon, oxide and metal, and their treatment, together with the processing of a very pure and defect-free (gate) oxide, made possible the production of devices which were no longer driven by a gate current but were distinguished by control using a gate voltage. In order to obtain such a possibility, the gate terminal has to be isolated with respect to the silicon by a thin oxide (a gate oxide about 0.1 μm thick), as schematically shown in Fig. 15.27. Again a low-n-doped layer (n^-) is responsible for the blocking capability. At the bottom in this figure a higher-doped n-region forms the drain region. With additional p-regions (p-wells) and n-regions on the upper side, an npn structure is obtained, similar to a bipolar transistor structure. But here the source terminal connects the upper p- and n-regions in order to suppress the influence of this pn junction. Whilst the n-source region is used to introduce electrons, the p-wells are necessary to establish a blocking capability and to act as a barrier against an unwanted current flow.

The isolated gate terminal is arranged in such a way that it overlaps the p-region. By means of a positive potential at the gate terminal with respect to the source terminal, electrons (negative) are pulled towards the silicon surface, creating a thin channel – an inversion layer – in the p-region connecting the n-source with the low-doped n^- middle region, opening the

p-barrier to allow a current flow. An accumulation layer is established in the n⁻ layer between the p-wells.

The concentration of electrons and therefore the conductivity of the channel can be regulated by the value of the positive gate potential. In order to obtain an n-channel in the p-region a minimum positive gate voltage, called the threshold voltage V_{th} , is necessary to enhance the electron concentration at the surface of the p-region until the p-concentration is about equalized. To obtain a high conductivity in the channel, the electron concentration has to be enhanced further until the p-doping concentration is distinctly exceeded. Then a path for an electric current carried by electrons is possible via the n-source, n-channel, n⁻-region and n-drain region to the drain terminal, characterized by an ohmic behavior. The various parts of the path have resistances $R_{nsource}$ (source), R_{ch} (channel), R_{n^-} (middle region) and R_{ndrain} (drain region), resulting in R_{Dson} (Fig. 15.27). The device behaves like a resistor that can be switched on and off.

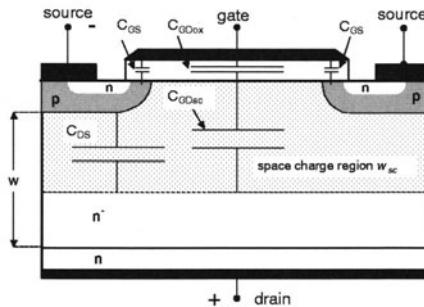


Fig. 15.28. The capacitances of an MOS transistor

Although one talks about a voltage-controlled device, it should be mentioned that the various capacitances present in the MOS transistor, shown in Fig. 15.28 for the blocking state, have to be charged and discharged, which requires a gate current but only for a short duration. The values of the capacitances depend on the thickness of the isolating gate oxide, the doping of the middle region, the channel area (the p-well area overlapped by the gate electrode) and the area between the p-wells.

If the polarity of the MOS transistor shown in Fig. 15.27 is changed to a positive potential at the source terminal and a negative potential at the drain terminal, the inherent but not optimized diode structure is activated and a swamping with carriers take place. This has to be considered if MOS transistors are used in circuits such as that shown in Fig. 15.4 where a change of the potentials at the switches occurs during the flow of the current through the freewheeling diodes.

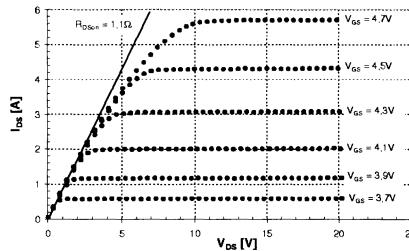


Fig. 15.29. Simulated V_{DS} - I_{DS} characteristics of a 500 V MOS transistor with an active chip area of 11 mm^2 at various gate voltages at 25°C

15.6.2 Conducting State

The simulated V_{DS} - I_{DS} characteristics in Fig. 15.29 show the dependence on the gate voltage for an MOS transistor with a blocking voltage of 500V and an active area of 11 mm^2 . At low drain-source voltages an ohmic region can be recognized. With increasing current at a fixed gate voltage V_{GS} , the voltage drop across the channel resistance R_{ch} reduces the effective gate voltage which is responsible for the electron concentration in the channel. With a lower electron concentration, a new balance occurs, resulting in a saturation of the current in the characteristic. This behavior is very advantageous, since in the case of a short circuit the current is limited and an electronic turn-off is possible. It should be mentioned that in such a case enhanced losses occur in the device, since the full circuit voltage and a current 5–10 times higher than the nominal current are present. When this situation occurs, the device has to be turned off within about $10 \mu\text{s}$.

For high-voltage devices ($V_{br} > 200 \text{ V}$), the low doping of the n^- -region dominates the value of the resistance R_{DSon} and therefore also the forward voltage drop V_{DSon} of the device. Since the n^- -region has to be chosen to support the blocking capability, for higher blocking voltages the doping n_D has to be lowered and, in addition, the dimension w of this region has to be extended. A twofold increase of the on-state resistance R_{DSon} results, according to

$$R = \frac{l}{\sigma \cdot A} = \frac{w}{q \cdot n_D \cdot \mu_n \cdot A}. \quad (15.22)$$

As schematically shown in Fig. 15.27, an effective area smaller than the chip area also has to be considered because no uniform current flow exists below the p-wells. The resistance is enhanced additionally by about a factor of 2 if the temperature is increased from 25°C to 125°C because the mobility μ_n of the electrons is reduced.

With a current of 10 A – a current density of $i \approx 1 \text{ A/mm}^2$ is typical for bipolar devices – for this 500 V MOS transistor, a voltage drop in the on-state of about 10 V would result, which is not tolerable for power devices since the losses can not be carried away by the cooling system.

In contrast, for blocking voltages $V_{br} < 200$ V very low R_{DSon} values are possible, making the MOS transistor favorable for low-voltage applications in the automotive field.

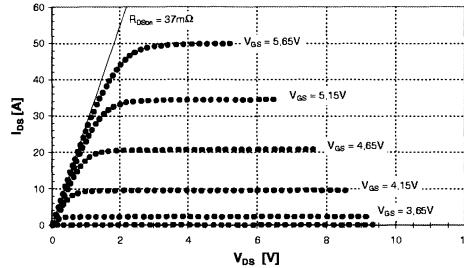


Fig. 15.30. Simulated V_{DS} – I_{DS} characteristics of a 100 V MOS transistor with an active chip area of 32 mm^2 at various gate voltages at 25°C

Typical V_{DS} – I_{DS} characteristics of a 100 V MOS transistor at 25°C with an area of 32 mm^2 are shown in Fig. 15.30. Now only a voltage drop of $V_{DSon} \approx 1$ V results at a current density of $i = 1 \text{ A/mm}^2$. For low-voltage MOS transistors, besides the doping of the n^- -region, other parts of the resistive path also contribute to R_{DSon} , e.g. the channel resistance R_{ch} , given by

$$R_{ch} = \frac{l_{ch}}{\sigma \cdot A_{ch}}. \quad (15.23)$$

Here, the channel cross section A_{ch} is given by

$$A_{ch} = w_{ch} \cdot d_{ch}, \quad (15.24)$$

where the channel length l_{ch} has to be kept short (a few microns) and the channel extension w_{ch} at the surface of the device has to be enlarged by suitable geometry. The depth d_{ch} of the channel is fixed at about 20 nm by physical laws, since the carriers are located just beneath the silicon surface and are held by the gate voltage.

The channel length l_{ch} has to be kept short (a few microns) and the channel extension w_{ch} at the surface of the device has to be enlarged by suitable geometry. The depth d_{ch} of the channel is fixed at about 20 nm by physical laws, since the carriers are located just beneath the silicon surface and are held by the gate voltage.

Using a cellular design shown in Fig. 15.31, with a pitch $a = 20 \mu\text{m}$ and the width of the p-well $b = 6 \mu\text{m}$, about 1 million cells can be placed in an area of 1 cm^2 . A channel width (the sum of the circumferences of all p-wells) of 240 m (!) results, and the channel resistance R_{ch} is reduced to a value below $0.5 \text{ m}\Omega$ for a device with a size of 1 cm^2 .

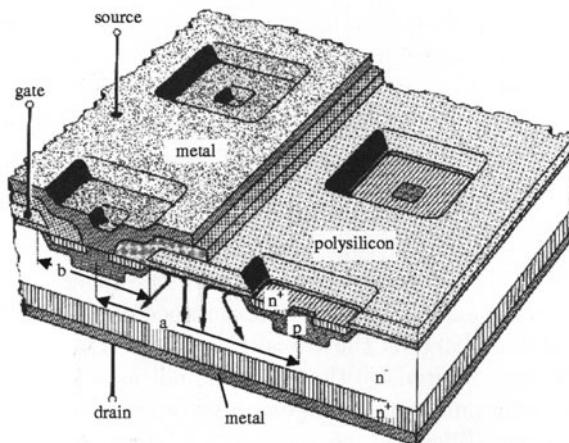


Fig. 15.31. Cell structure of a power MOS transistor

15.6.3 Dynamic Behavior

The capacitances inherent in the MOS transistor, as shown in Fig. 15.28, have to be charged and discharged during the switching process. The values of these capacitances are partly dependent on the voltage across the device and are not constant during the switching process.

Turn-On

Typical waveforms for the drain–source current I_{DS} , the drain–source voltage V_{DS} and the gate voltage V_{GS} during the turn-on of an MOS transistor are shown in Fig. 15.32; these waveforms were obtained by simulation.

The turn-on process usually starts from a blocking condition of the device; in this condition the supply voltage V_{CC} (400 V in this figure) has been established across the device. A turn-on of the device can be initiated by

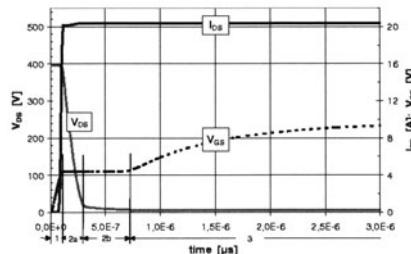


Fig. 15.32. Waveforms of voltage V_{DS} , current I_{DS} and gate voltage V_{GS} during the turn-on of an MOS transistor

changing the supply voltage of the gate circuit V_G from zero (or from a negative value) to a positive potential, e.g. $V_G = 10\text{ V}$.

The capacitances C_{GS} and C_{GD} (see Fig. 15.28) have to be charged until a positive voltage high enough to produce a channel across the p-wells occurs at the gate. Since a blocking voltage is present across the device, causing a space charge region of about $100\text{ }\mu\text{m}$, the capacitance of the space charge C_{GDsc} (see Fig. 15.28) is small ($\approx 0.1\text{ nF}$ for 1 cm^2). The capacitance C_{GS} ($\approx 5\text{ nF}$ for 1 cm^2) dominates (time phase 1 in Fig. 15.32) because the large oxide capacitance C_{GDox} and the capacitance of the space charge C_{GDsc} are in a series connection and the smallest capacitance dominates C_{GD} .

If the gate voltage exceeds the threshold voltage V_{th} , the flow of a current starts until it is limited by the circuit conditions (e.g. $I_{DS} \approx 20\text{ A}$ in Fig. 15.32). The minimum gate voltage necessary for this current flow in the channel is called the Miller voltage $V_{Gmiller}$ and corresponds to the stationary voltage-current characteristics; no further charging of the gate would be needed.

But the flow of the gate current I_G continues if the gate supply voltage V_G is higher than the Miller voltage $V_{Gmiller}$ at the device, according to

$$I_G = \frac{(V_G - V_{Gmiller})}{R_G}, \quad (15.25)$$

and a reduction of the space charge region results because additional electrons can neutralize positive donors in this region. A reduction of the space charge region is synonymous with a decreasing voltage across the device. Further charging of the gate is delayed because the gate current is now used to provide a change of the voltage dV/dt according to

$$I_G = C_{GD}(V) \cdot \frac{dV}{dt}. \quad (15.26)$$

A constant voltage at the gate – the Miller voltage $V_{Gmiller}$ – is the result, with a nearly constant decrease of the drain-source voltage V_{DS} down to a value of about 20 V . The gate current determines the rate of the voltage decrease (time phase 2a in Fig. 15.32) in accordance with (15.26).

At voltages $< 20\text{ V}$ dV/dt is slowed down, and also the extension of the space charge zone shrinks to a few microns and the value of the capacitance C_{GD} starts to grow. The gate voltage at the MOS transistor is still constant because now the gate current is used to charge the increasing capacitance (phase 2b in Fig. 15.32):

$$I_G = \frac{d(C_{GD}(V))}{dt} \cdot V(t). \quad (15.27)$$

An ongoing charging of the gate with a gate current

$$I_G = \frac{(V_G - V_{GS}(t))}{R_G} \quad (15.28)$$

up to the value of the gate supply voltage (e.g. $V_G = 10$ V) is only possible after the time-dependent changes of the voltage and the capacitances have been finished (time phase 3 in Fig. 15.32).

But in this phase 3, the capacitances ($C_{GS} + C_{GD}$) are in the order of 100 nF because the inversion layer in the p-well and the accumulation layer in the n⁻-region exist and only the thickness of the gate oxide (about 0.1 μm) determines the capacitance, therefore enlarging the value by a factor of about 10–100. The further charging time is prolonged considerably, as shown in Fig. 15.32, phase 3.

At the beginning of phase 3, according to (15.25), the charging of the capacitances depends on the difference between the gate supply voltage V_G and the Miller voltage $V_{Gmiller}$. The Miller voltage itself is determined by the doping concentration of the p-wells. If the Miller voltage is closer to the applied gate supply voltage, the charging time in this phase is enlarged additionally.

Turn-Off

Typical waveforms for I_{DS} , V_{DS} and V_{GS} during a turn-off are shown in Fig. 15.33. The starting condition was a device in the conducting state with a positive voltage of $V_{GE} = 10$ V at the gate terminal with respect to the source terminal and with a current flow made possible by the channel in the p-region. The turn-off process can be initiated if the gate supply voltage is changed to zero or a negative value. A gate current flows from the positive potential at the gate of the device, established during the conducting state, to the negative potential of the gate voltage source and discharges the gate until the minimum gate voltage – the Miller voltage $V_{Gmiller}$ – necessary for a continuing current flow is reached. Although capacitances identical to those in Fig. 15.32, phase 3, have to be discharged, the discharging time is now shorter because a larger gate current is flowing. This is made possible by a larger voltage difference between the positive gate voltage V_{GE} (10 V and later the positive Miller-voltage $V_{Gmiller} \approx 7$ V in this case) and the gate supply voltage V_G in the gate circuit (0 V or even a negative value).

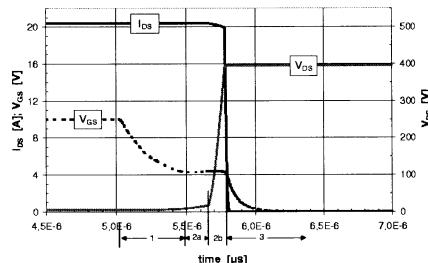


Fig. 15.33. Waveforms of voltage V_{DS} , current I_{DS} and gate voltage V_{GE} during the turn-off of an MOS transistor

Now the gate current has a direction opposite to the drain–source current and reduces the (electron) current in the channel; it therefore also lowers the number of electrons delivered by the channel (phase 1 in Fig. 15.33). Since a continuing constant current is flowing more electrons are leaving the middle region at the drain terminal than are delivered by the channel. The missing electrons are substituted first from the accumulation layer just below the surface of the n^- -region and later from the low-doped middle region, leaving behind uncompensated donors in this region.

At the beginning of this process, the electrons removed from the accumulation layer at the surface of the n^- -region and the start of the depletion of the middle region lower the capacitance C_{GD} distinctly because the relevant distance is changed from $0.1\text{ }\mu\text{m}$ – the gate oxide thickness – to the thickness of a space charge zone of a few microns. The gate current is used to discharge the decreasing capacitance, and further discharging of the gate is delayed (phase 2a in Fig. 15.33).

With a growing space charge zone, the change in the capacitance is reduced but a voltage is built up across the device. Again the gate current determines the voltage increase, in a similar way to the case of turn-on (phase 2b in Fig. 15.33), in accordance with (15.26).

These discharging processes are not finished before the supply voltage V_{CC} is reached. After this time instant, the gate current discharges the gate very quickly to the value given by the gate supply voltage (phase 3 in Fig. 15.33).

15.6.4 Trends

As discussed for the low-voltage MOS transistor, all parts contributing to R_{DSon} have to be minimized. Besides efforts to minimize the channel resistance, also introducing the trench technology well known for memory devices can distinctly lower the resistance R_{DSon} by making possible the reduction of the p-well areas and reduced distances between the p-wells [32, 33].

However, the strong correlation between the blocking-voltage capability and the conductivity in the conducting state has limited the application of MOS transistors in power applications more or less to voltages smaller than 100 V. Efforts to make a breakthrough between these two contradicting demands have created a structure using high-doped columns alternately doped n and p. The doping of the columns has to be chosen in such a way that the resulting net doping tends to zero. Such a structure is available, e.g. a structure from Infineon Technologies called Cool-MOSTM, and is described in Chap. 16 in this book.

This invention has extended the usage of MOS transistors to blocking voltages up to about 800 V in order to take advantage of the fast switching capabilities and low switching losses of these devices, allowing switching frequencies in the range of 100 kHz as preferred in switch mode power supplies.

15.7 IGBT (Insulated-Gate Bipolar Transistor)

15.7.1 Behavior in Principle

At the beginning of the 1980s a device structure quite similar to the MOS transistor was proposed; the only difference was that the n-region at the bottom (drain) in Fig. 15.27 was changed to a p-region (Fig. 15.34). Because no difference exists in principle at the upper side, the function of this device is comparable to the situation for the MOS transistor. In this case also, a channel has to be formed by electrons under the influence of a positive gate potential to enable a flow of an electron current. These electrons flow towards the p-region at the bottom, which forms a pn junction with the n⁻-region. This pn junction is biased in the forward direction. Although holes are emitted from this p-region, it has been called the “collector” following the terminology of the bipolar transistor.

In the low-doped n⁻-region, an electron-hole plasma similar to those in the diode and thyristor can be established. The shape of the carrier distribution can be adjusted by the geometry at the emitter side and by the emitter efficiency at the collector side. The doping of the n⁻-region does not play any role in the forward voltage drop, in contrast to the situation in the MOS transistor (shown for comparison in Fig. 15.35).

The neutrality in the n⁻-region is responsible for the fact that the electrons and the holes flow along the same path for as long as possible, and a split-up of the current occurs near the channel (Fig. 15.34). Whereas the electrons use the channel path, the holes flow in the p-region just beneath the channel towards the emitter terminal, bypassing the n-region.

Since usually this upper p-region (p-well) has only a moderate conductivity, a voltage drop originates from the hole current flowing in this region (indicated in Fig. 15.34), and an internal forward bias of the existing pn junction occurs, although both regions are connected by the emitter electrode.

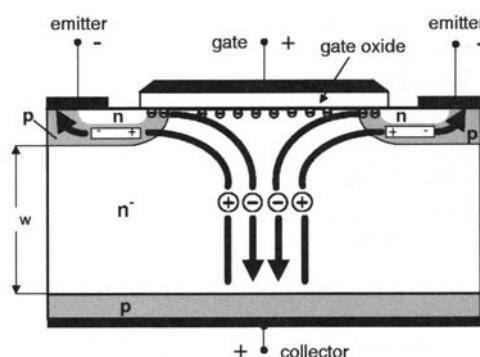


Fig. 15.34. Cross section of an IGBT with the flow of electrons and holes in the conducting state

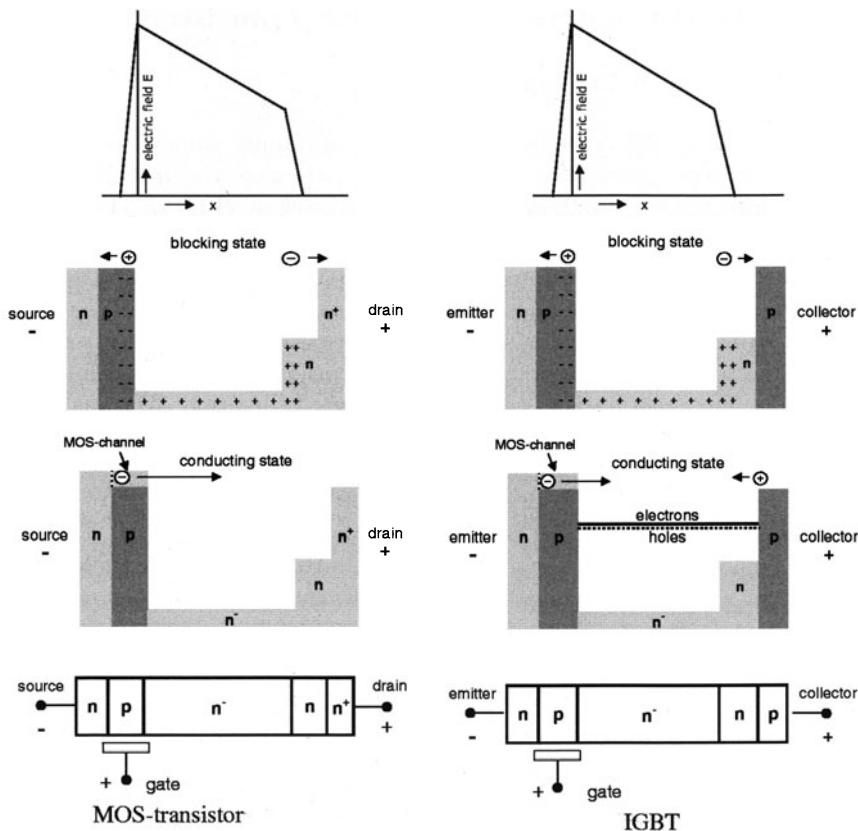


Fig. 15.35. Comparison between MOS transistor and IGBT in the blocking and conducting state

The additional electrons can act as a trigger current for the inherent npnp structure, similarly to the situation in a thyristor structure. Because a loss of the control mechanism of the gate would result, such an unwanted behavior has to be avoided by reducing the critical dimensions to a few microns. In addition, the distances between the upper p-wells have to be designed carefully to obtain carrier concentrations in the middle region that are optimal for a low forward voltage drop V_{CEsat} .

An activation of the upper pn junction is also possible in the MOS transistor structure. The npn structure of the MOS transistor has no regenerative behavior like that of an npnp thyristor structure, but a loss of gate control is possible. Since the current densities are usually distinctly higher for IGBTs, this behavior requires more attention in the design of the IGBT structure.

In the case of a reversal of the potential on the device, in contrast to the behavior of an MOS transistor, the pn junction at the collector side in the

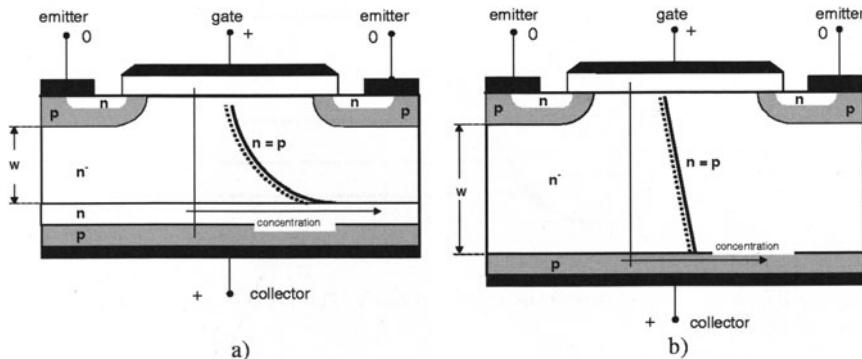


Fig. 15.36. PT (punch-through) concept (a) and NPT (non-punch-through) concept (b), with the carrier distribution in the middle region in the conducting state

IGBT structure shows a blocking characteristic – although this behavior is not cultivated – which avoids the activation of a diode behavior.

As shown in Fig. 15.36, two structures have been developed for IGBTs. In Fig. 15.36a the **punch-through** concept (PT-IGBT), with an n-buffer layer in front of the high-doped p-collector layer is shown, which possesses high emitter efficiency [9–12]. In order to obtain an acceptable switching behavior, a lifetime doping is necessary. For the **nonpunch-through** concept (NPT-IGBT, shown in Fig. 15.36b), the width of the low-doped n-type middle region has to be large enough to avoid a punch-through of the electric field [34]. The height of the carrier concentration is adjusted by a low emitter efficiency of the pn collector junction and a lifetime doping is not necessary. The carrier distributions typical of PT- and NPT-IGBTs are also shown in this figure.

In the past, the NPT structure has shown the more rugged behavior because a change of the carrier lifetime caused by an elevated temperature has a less significant influence on the carrier distribution and the reduced carrier mobility provides an increasing forward voltage drop V_{CEsat} , stabilizing the current distribution if chips are operated in parallel (see Fig. 15.38).

15.7.2 Conducting State

The $V_{CE}-I_{CE}$ characteristics of the IGBT are comparable in their shape to those of the MOS transistor if the dependence on the gate voltage is considered (Fig. 15.37). But with the additional pn junction at the collector side, which is biased in a forward direction, the characteristics of the IGBT behave like a diode characteristic, with a threshold voltage and a continuous improvement of the conductivity with increasing current density due to the swamped middle region (Fig. 15.38). The characteristics shown in Figs. 15.37 and 15.38 belong to a 1200 V/1200 A IGBT in a modular construction as shown in Fig. 15.3, where 24 IGBT chips are arranged in parallel (in addition, 12 freewheeling diodes are placed in such a module).

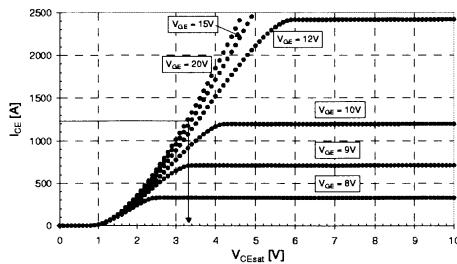


Fig. 15.37. V_{CE} – I_{CE} characteristics of a 1200 V/1200 A IGBT module at various gate voltages V_{GE}

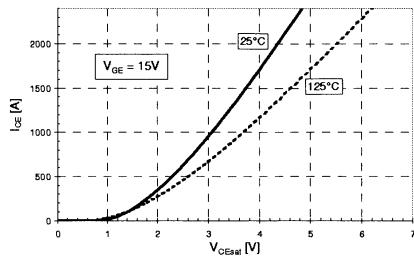


Fig. 15.38. V_{CEsat} – I_{CE} characteristics of a 1200 V/1200 A IGBT module at 25°C and 125°C

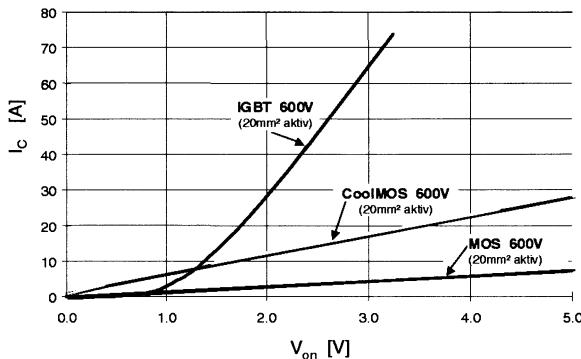


Fig. 15.39. Comparison between 600 V devices (MOS transistor, Cool-MOS™ and IGBT) depicting the forward voltage drop V_{on} in the conducting state

The saturation in the V – I characteristic in Fig. 15.37 has the same reason as discussed for the MOS transistor, but now only a part of the load current – the electron current, which is about $3/4$ of the load current – reduces the effective gate voltage.

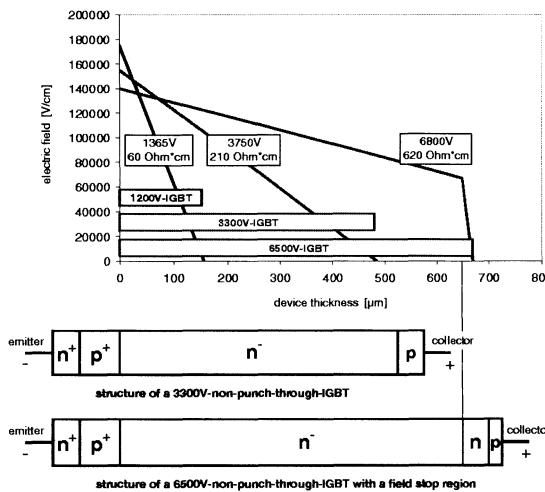


Fig. 15.40. Electric field, chip thickness and doping of the middle region of IGBTs with various blocking-voltage capabilities

In Fig. 15.39, the forward voltage drops of an MOS transistor, a Cool-MOSTM and an IGBT are compared, all devices having a 600 V blocking capability and also the same active chip area (20 mm^2). The advantage of the IGBT is convincing, and the Cool-MOSTM structure is only superior in a low-current range. For higher blocking voltages, the advantage of the IGBT is still more pronounced.

15.7.3 Blocking-Voltage Capability

Since a pnp structure is present, the precautions concerning punch-through of the electric field have to be considered, as discussed for thyristor structures. With an only moderately more highly doped field stop layer in front of the p-collector layer, the advantage of a reduced width of the middle region, as in the PT concept, can be utilized without losing the advantages of the NPT structure.

In Fig. 15.40, the field distributions in the low-doped middle region, the blocking-voltage capabilities, the chip thickness required and the doping values for the starting silicon material are shown for IGBTs realized today.

15.7.4 Dynamic Behavior

For the switching behavior, the same capacitances as discussed for the MOS transistor participate in the switching process (Fig. 15.41). Only the values of the capacitances are different, owing to different design rules used for MOS transistors and IGBTs. The waveforms of the voltage V_{CE} , current I_C and

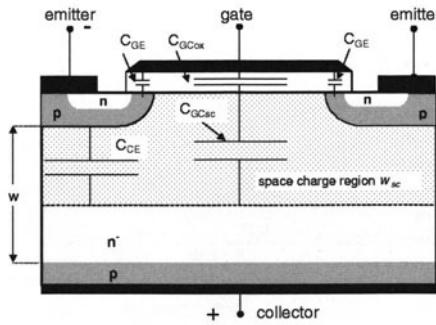


Fig. 15.41. The capacitances of an IGBT in the blocking state

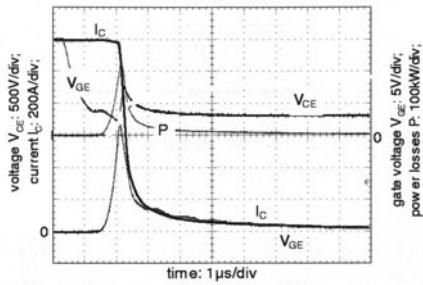


Fig. 15.42. Measured waveforms of voltage V_{CE} , current I_C and gate voltage V_{GE} during turn-off of a 3300 V/1200 A IGBT module

gate voltage V_{GE} during turn-off and turn-on of the IGBT are similar to those for the MOS transistor. Again the gate current controls the increase and decrease of the voltage across the IGBT during turn-off and turn-on, respectively.

If, during turn-off, the voltage across the device is equal to the circuit supply voltage V_{CC} , a rapid decrease of the current follows. Since the current densities are higher than in the MOS transistor, higher voltage spikes can appear, especially if high-current switches are used with IGBT chips arranged in parallel.

In Fig. 15.42, the turn-off behavior of a 3300 V/1200 A IGBT module is shown. When turn-off is initiated by a negative gate supply voltage, the gate of the IGBT is discharged, reducing the gate voltage. After about 2 μ s, the current decreases within less than 1 μ s from 1200 A to 200 A, but for the next few microseconds a much slower decrease follows, the tail current of the IGBT flowing for several microseconds.

In contrast to the behavior of the MOS transistor in Fig. 15.32, the IGBT shows a current tail during turn-off, which is caused by the charge still stored after the voltage is built up. The tail current effects the removal of this remaining charge. This turn-off behavior of the IGBT results in higher switch-

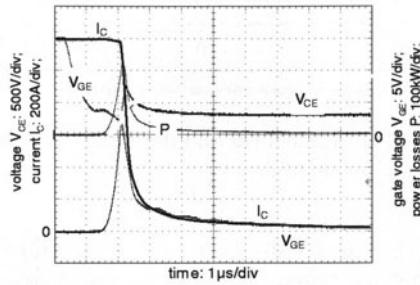


Fig. 15.43. Measured waveforms of voltage V_{CE} , current I_C and gate voltage V_{GE} during turn-on of a 3300 V/1200 A IGBT module

ing losses, which are significant if switching frequencies higher than 10 kHz are used and are necessary.

In Fig. 15.43, the measured turn-on behavior of a 3300 V/1200 A IGBT module is shown under rated conditions, where the device is turned on from a blocking voltage of 1800 V to a current level of 1200 A. After a delay time of about 2 μ s, the gate current has charged the gate to the Miller voltage of about 11 V and the increase of the load current starts. The current increase and the voltage decrease take place within 1 μ s. The final charging of the gate continues for 4 μ s after the beginning of the current increase. In the current waveform, an overshoot to 2240 A is visible; this is the reverse recovery current of the freewheeling diode, as discussed in connection with Fig. 15.4. The peak losses occurring during these switching phases are 350 kW (turn-on) and 220 kW (turn-off).

15.7.5 Trends

Today the IGBT is in widespread use, replacing bipolar transistors in the low- and medium-power range, and is also starting to advance into the high-power field, still dominated by thyristors and GTOs today. The progress in the chip-shrinking of the NPT-IGBT achieved by introducing trench technology and a field stop structure is shown in Fig. 15.44 for a 1200 V/75 A IGBT chip; this has resulted in distinctly reduced manufacturing costs [35]. Today such a chip consumes less than half the area needed 10 years ago, and has a lower V_{CEsat} and reduced switching losses as well.

Although some years ago a limit for the IGBT was postulated at blocking voltages of about 2000 V, there is no physical limitation in principle. But in high-voltage applications the requirements on the device have to be reduced, taking account of the current density and the value of dV/dt during the switching process as has to be done for other devices also. Since an easy control is possible, e.g. by a resistor in the gate circuit, switching in high-voltage circuits is also possible.

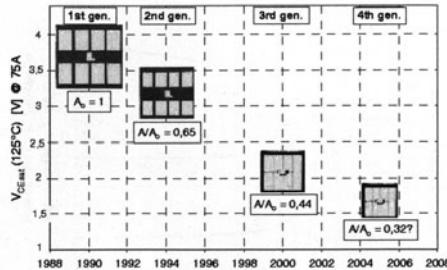


Fig. 15.44. The progress in the development of IGBTs, considering V_{CESat} and the chip size for a 1200 V/75 A IGBT chip [35]

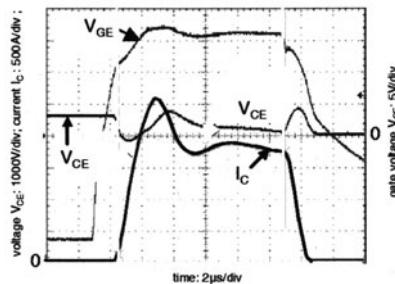


Fig. 15.45. Waveforms of voltage V_{CE} , current I_C and gate voltage V_{GE} during a short-circuit condition

As an example of the ruggedness of the IGBT, Fig. 15.45 shows the short-circuit capability of a 6500 V/600 A IGBT module [36]. With a supply voltage of 4500 V, a short-circuit current of 2700 A is the result, but the IGBT can be switched off after 8 μ s without destruction of the device. Whereas 6500 V/600 A IGBT modules are being manufactured successfully today, a further extension of the blocking capability has been reported in [37].

15.8 Conclusions

As mentioned in the introduction, the development of silicon power semiconductor devices requires the existence of a suitable material. In the 1950s this material had to be developed and manufactured in parallel with the devices themselves. The first important steps in the development of the material silicon were the fabrication of monocrystalline silicon containing an impurity content as low as possible. Also, the manufacturing of silicon rods with large diameters and a defined doping of this silicon, necessary for npn and npnp structures, was a problem that had to be solved.

For devices with high blocking voltages and large base widths, the carrier lifetime has to be as high as possible during device production in order either

to obtain a low forward voltage drop or to make possible a well-defined reduction of the carrier lifetime, as necessary for the dynamic behavior of the device.

In the past 50 years many power semiconductor devices were proposed, but only a few of these – described in this chapter – are actually produced on a large scale and are used in the field of power electronics in large numbers today.

The driving force behind the development of integrated circuits towards continuously higher integration does not apply to bipolar devices, because all the volume is swamped by carriers and the functions of additional structures on the same chip suffer under this condition. Sometimes the integration of a switch together with its antiparallel diode has been realized, but often such integration fails because the requirements for these two devices are too different.

Integration is possible for MOS power devices (e.g. together with an MOS transistor), which have a unipolar conductivity and are usually not swamped by carriers. Such devices are used in automotive applications in the low-voltage range, but it has to be taken into account that additional functions often result in more complex processing steps, making the manufacturing costs more expensive.

References

1. J. Bardeen, W.H. Brattain: The transistor, a semiconductor triode. *Phys. Rev.* **74**, 230 (1948)
2. W. Shockley: The theory of p–n junctions in semiconductors and p–n junction transistors. *Bell Syst. Tech. J.* **28**, 435 (1949)
3. R.N. Hall: Power rectifiers and transistors. *Proc. IRE* **40**, 1512 (1952)
4. J.L. Moll, M. Tanenbaum, J.M. Goldey, N. Holonyak: PNPN-transistor switches. *Proc. IRE* **44**, 1174 (1956)
5. T. Nagano, M. Okamura, T. Ogawa: A high-power, low-forward-drop gate turn-off thyristor. *Conf. Rec. IEEE-IAS*, 1003 (1978)
6. M. Azuma, A. Nakagawa, K. Takigama: High power gate turn-off thyristors. *Japan. J. Appl. Phys.* **17-1**, 275 (1978)
7. J. Tihanyi, P. Huber, J.P. Stengl: Switching performance of vertical Siemens power MOSFETs. *IEDM Tech. Digest*, 692 (1979)
8. V.A.K. Temple, P. V. Gray: Theoretical comparison of DMOS and VMOS structures for voltage and on-resistance. *IEDM Tech. Digest*, 88 (1979)
9. H.W. Becke, C.F. Wheatley: Power MOSFET with an anode region. U.S. Patent 4,634,073, issued Dec. 1982
10. J.P. Russell, A.M. Goodman, L.A. Goodman, J.M. Neilson: The COMFET – a new high conductance MOS-gated device. *IEEE Electron Device Lett.* **EDL-4**, 63 (1983)
11. B.J. Baliga, M.S. Adler, P.V. Gray, R.P. Love, N. Zommer: The insulated gate rectifier (IGR): a new power switching device. *IEDM Tech. Digest*, 264 (1982)
12. A. Nakagawa, T. Tsukakoshi, H. Ohashi: High voltage bipolar-mode MOSFET's with high current capability. *Ext. Abstr. 16th Conf. Solid State Devices and Materials* (1984) pp. 309–312

13. S. Sze: *Physics of Semiconductor Devices* (Wiley, New York 1981)
14. H. Benda, E. Spenke: Reverse recovery processes in silicon power rectifiers. Proc. IEEE **55**, 1331 (1967)
15. G. Deboy, G. Sölkner, E. Wolfgang, W. Claeys: Absolute measurement of transient carrier concentration and temperature gradients in power semiconductor devices by internal IR-laser deflection. Microelectron Eng. **31**, 299 (1996)
16. A. Herlet: Maximum blocking capability of silicon thyristors. Solid State Electron. **8**, 655 (1965)
17. M. Schnöller: Breakdown behavior of rectifiers and thyristors made from striation-free silicon. IEEE Trans. Electron Devices **ED-21**, 313 (1974)
18. K. Platzöder, K. Loch: High-voltage thyristors and diodes made of neutron-irradiated silicon. IEEE Trans. Electron Devices **ED-23**, 805 (1976)
19. A. Herlet, K. Raithel: Forward characteristics of thyristors in the fired state. Solid State Electronics **9**, 1089 (1966)
20. W.H. Dodson, R.L. Longini: Probed determination of turn-on spread of large area thyristors. IEEE Trans. Electron Devices **ED-13**, 478 (1966)
21. I. Somos, D.E. Piccone: Some observations of static and dynamic plasma spread in conventional and new power thyristors. IEE Conf. Rec. **53**, 1 (1969)
22. F.E. Gentry, J. Moyson: The amplifying gate thyristor. Proc. IEDM (1968) p. 110
23. A. Herlet, P. Voss: State of the art in power semiconductor design. Conf. Rec. IEEE-IAS Int. Semicond. Power Conv. Conf. (1977) pp. 7–23
24. P. Leturcq: Power devices: specific problems. ESSDERC Conf. Rec. (1975) pp. 119–153
25. H.-J. Schulze, M. Ruff, B. Baur: Light triggered 8kV thyristor with a novel integrated breakdown diode. Proc. ISPSD (1996) p. 197
26. H. Gruening, B. Oedegard, A. Weber, E. Carroll, S. Eicher: High-power hard-driven GTO Module for 4,5kV/3kA snubberless operation. Proc. PCIM (1996) pp. 169–183
27. H.A. Schafft: Second breakdown – a comprehensive review. Proc. IEEE **55**, 1272 (1967)
28. P.L. Hower: Optimum design of power transistor switches. IEEE Trans. Electron Devices **ED-20**, 426 (1973)
29. C. Hu, M.J. Model: A model of power transistor turn-off dynamic. IEEE Power Electronics Specialists Conf. (1980) pp. 91–96
30. P.L. Hower: A model for turn-off in bipolar transistors. IEDM Tech. Digest (1980) pp. 289–292
31. G. Miller, A. Porst, H. Strack: An advanced high voltage bipolar power transistor with extended RBSOA using 5µm small emitter structures. Proc. IEDM (1985) pp. 142–145
32. D. Ueda, T. Takagi, G. Kano: A new vertical power MOS-FET structure with extremely reduced on-resistance. IEEE Trans. Electron Devices **ED-32**, 2 (1985)
33. H.R. Chang, B.J. Baliga, J.W. Kretchmer, P.A. Piacente: Insulated gate bipolar transistor (IGBT) with a trench gate structure. Proc. IEDM (1987) pp. 674–677
34. G. Miller, J. Sack: A new concept for non-punch-through IGBT with MOSFET-like switching characteristics. IEEE Power Electronics Specialists Conf. (1989) pp. 21–25

35. T. Laska, G. Miller, M. Pfaffenlehner, P. Türkes, D. Berger, B. Gutzmann, P. Kanschat, M. Münzer: Short circuit properties of trench-/field-stop-IGBTs – design aspects for a superior robustness. Proc. ISPSD (2003) pp. 152–155
36. J.G. Bauer, F. Auerbach, A. Porst, R. Roth, H. Ruething, O. Schilling: 6,5 kV-modules using IGBTs with field stop technology. Proc. ISPSD (2001) pp. 121–124
37. M. Rahimo, A. Kopta, S. Eicher, N. Kaminski, F. Bauer, U. Schlapbach, S. Linder: Extending the boundary limits of high voltage IGBTs and diodes to above 8 kV. Proc. ISPSD (2002) pp. 41–44

16 Compensation Devices Break the Limit Line of Silicon

G. Debay

16.1 Introduction

Today's lifestyle is no longer conceivable without the generation, distribution and conversion of electrical energy. We daily use numerous electrically driven machines, which draw their energy either from AC or DC rails or from batteries. All of these applications require a power conversion from the input line to the desired output voltage. In the past this task was often accomplished by linear voltage regulators, which in the case of AC/DC conversion used huge, heavy chokes. In times of mobile communication and global business, where products are no longer developed for local markets but rather are sold worldwide, this approach is no longer satisfactory. From the manufacturers point of view, platform strategies have the advantage of reduced production complexity and better purchasing conditions due to increased volume. The other driving force is the mobile use of equipment in countries with different AC power supplies: a mobile phone battery may be refueled today in Germany, tomorrow in the US and next week in Japan. The charger has to supply in all cases the same output voltage and current from substantially different input lines. This requirement is best fulfilled today by switch mode power supplies (SMPS). These power converters work typically in the range of 70 to 100 kHz instead of 50 Hz and may therefore use a small, lightweight transformer. This difference in weight is obvious when a switch mode power supply is compared with a linear regulator. A key component of an SMPS is the high-voltage switch, which typically sustains a blocking voltage in the range of 500 to 800 V depending on the chosen circuit topology. A widespread topology for the low end of the power range of up to 200 W is the flyback converter, which transforms the energy in the blocking phase of the switch. This topology requires transistors with 500 to 600 V blocking capability.

The main driving factor for the huge and fast-growing market segment of AC/DC converters is cost, which means in the first place the product cost of the high-voltage switch, but also system costs, which may be retranslated into performance aspects of the transistor. As an example, there is the need for low driving power/gate charge, which saves in the cost of the control IC; high efficiency, which means no or only a small heat sink; small packages, which save space on the printed circuit; and last but not least, low EMI noise, which helps to save costs in the filter section. Another important aspect which is

helping to substitute linear regulators with switch mode power supplies is legal requirements on the power factor. The demanding price pressure from the system and device levels is forcing the manufacturers of semiconductor devices to look for innovative solutions to meet tomorrow's market requirements with competitive products.

16.2 Today's High-Voltage Device Concepts and the Way Towards the Compensation Principle

The challenge of the high-voltage switch is to master simultaneously the need for a high blocking voltage and high forward currents at low conduction losses. Furthermore the switch should be easily controlled, allow fast switching and should not die under mild overload conditions, and all that at low cost. If we exclude mechanical relays from our approach (which would show a lifetime of only a few minutes if switched in the kHz range), our view is focused on semiconductor solutions. With regard to design, the task of high blocking voltage means low doping and a large thickness of the voltage-sustaining region, whereas low conduction losses require just the opposite. This fundamental conflict was historically first solved by bipolar transistors, where the low-doped voltage-sustaining region is conductivity modulated by several orders of magnitude by the injection of a highly conductive electron-hole plasma. Unfortunately these devices require in the on-state a continuous base current which is only one or two orders of magnitude smaller than the load current, thus adding complexity to the driving circuit. In conjunction with relatively large parameter variations, this disadvantage means that these devices are today only found in niche applications such as line deflection in colour TV (CTV) sets and absolutely low-price segments such as the electronic ballast of cold fluorescent lamps (CFLs).

The combination of bipolar conductivity modulation with MOS gated current control led to the insulated-gate bipolar transistor or, for short, IGBT. Instead of controlling the transistor via a base current, the steering electrons flow through an MOS channel, which may be formed laterally or vertically in the transistor cell. The voltage-sustaining region may also be arranged laterally (parallel to the surface of the device) or vertically (from the front to the back of the device). When this electron current reaches the rear side of the device, it forward-biases the pn junction there, leading to an injection of holes. This inherent diode structure leads to a minimum forward voltage drop of 0.7 V (at least in silicon) for all IGBT concepts.

Because the neutrality condition, the local numbers of electrons, holes and ionized impurities remain unaltered. This ambipolar balance of carriers is one of the secrets of stability of the IGBT. According to the electric-field profile, we distinguish between two major concepts, the **punch-through**, or PT, concept and the **nonpunch-through**, or NPT, concept. Figure 16.1 shows

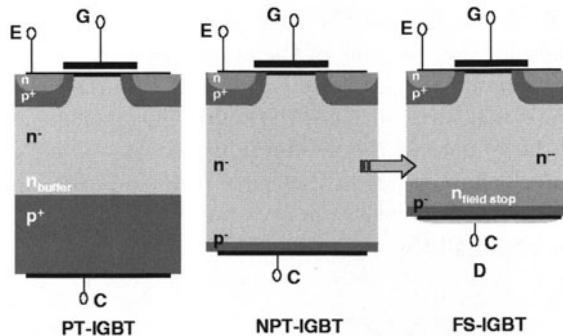


Fig. 16.1. Cross sections of the PT-IGBT and NPT-IGBT, and a further development step called the FS-IGBT

both device types – here with a lateral MOS channel and vertical drift region – in cross section.

The PT concept was first proposed in the early 1980s, whereas the NPT concept was developed in the late 1980s. The first of these device concepts has owing to its trapezoidal field profile in the blocking state, the advantage of a relatively thin voltage-sustaining layer. The field is stopped in the high-doped n buffer. In the on-state, a thin layer to be flooded with electron–hole plasma means less charge to be removed during turn-off. The main advantage of the PT-IGBT is therefore its superior trade-off between conduction and switching losses. There are, however, two disadvantages: on the one hand, owing to the high emitter efficiency of the thick back p⁺-layer, there is the need to reduce the free-carrier lifetime in the active n⁻-doped area to make the PT-IGBT an appropriate device for high frequencies and therefore suitable for SMPS applications. The techniques used here involve the creation of mid-bandgap centers by electron or heavy-ion-irradiation or diffusion of Pt, Au etc. These recombination centers, unfortunately become rather ineffective at elevated device temperatures leading to an undesired increase in turn-off losses with rising temperature. Furthermore, the PT concept is inherently based on epitaxial wafer material, which is a major cost issue.

Here lies the main advantage of the NPT concept, which requires wafers with only one doping type and concentration for the voltage-sustaining region. In the case of an n-doped substrate material, the p-emitter on the back is implanted and annealed, and its efficiency may therefore easily be controlled. The trade-off between conduction and turn-off losses is, however, inferior to that in the PT concept owing to the triangular field profile, which requires a thicker active region to sustain the same blocking voltage. Even at the maximum voltage the space charge region does not reach the back-surface p-emitter.

The ideal solution would therefore be to combine the low-cost base material and the shallow emitter of the NPT-IGBT with the trapezoidal field of

the PT-IGBT. And that is what the field stop IGBT, or FS-IGBT, is about. This concept therefore shows none of the above-mentioned disadvantages but all of the major benefits [1]. As there is, however, no free lunch, the concept requires elaborate measures in the production line to cope with ultra-thin wafers during a large part of the production process [2]. For a 600 V IGBT, the required wafer “thinness” is in the range of 45 to 70 μm , which resembles more a razor blade than the usual massive wafer thickness. Figure 16.2 shows the trade-off curves of all the concepts for 600 V devices.

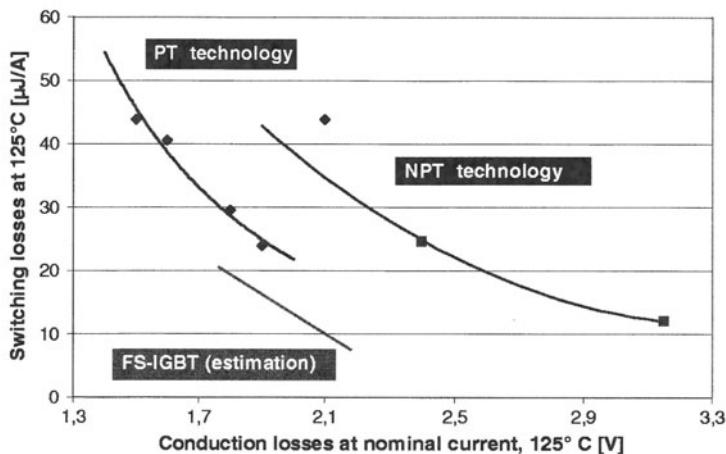


Fig. 16.2. Comparison of today's state of the art for 600 V IGBTs: switching losses versus conduction losses

IGBTs have today conquered a large application field, ranging from low-frequency drives, for example in washing machines and large-scale traction applications, up to high-frequency applications in switch mode power supplies. In high-frequency applications, however, the switching losses of the IGBT may become an obstacle. Another aspect is the threshold of the inherent diode structure, which leads at minimum to conduction power losses of around 1 W per ampere.

In these cases unipolar device concepts are the right choice. Here, the current is conducted by one carrier type only, e.g. electrons. There is no conductivity modulation of the voltage-sustaining region. The main representative of this device concept is the power MOSFET, which was developed back in the mid 1970s to the early 1980s. The current is controlled via a lateral or vertical MOS channel structure, with the voltage-sustaining drift region being also laterally or vertically oriented. For a high blocking voltage a low doping level of the drift region is required. The electric field has its maximum at the blocking pn junction and decreases from that position in proportion to the space charge as described by Poisson's law. Owing to the fact that

there is no conductivity modulation of the drift region by injected carriers, the specific resistance (per unit area) is relatively high. It increases as a function of voltage with an exponent of roughly 2.5. This factor derives from the necessity to increase the thickness of the voltage-sustaining region and lower its doping level simultaneously as the voltage is increased. Furthermore, the critical electric field for breakdown decreases at lower doping concentrations.

This major disadvantage has favored application topologies which are able to cope with a lower breakdown voltage of the device, e.g. the half bridge using 500 V transistors that is used as a lamp ballast.

Continuous improvement of the power MOSFET in the 1980s and 1990s could not solve this fundamental problem. A doping profile rising within the voltage-sustaining region for better RDson of the drift region or increased doping between the p-wells for minimizing this inherent JFET structure has been proposed [3]. The most recent development and today's state of the art uses very narrow n-regions between the p-wells and great effort in the construction of the edges of the device in order to come as close as possible to the breakdown voltage of a one-dimensional, one-sided, abrupt pn junction. The group that has performed this development achieves 97% of the maximum breakdown voltage of a given structure and therefore reaches 110% of the so-called silicon limit [4]. This limit is the mathematical solution of the problem of optimizing RDson and the blocking voltage of an abrupt p^+n^- junction with one n-doping profile. Figure 16.3 shows the corresponding designs.

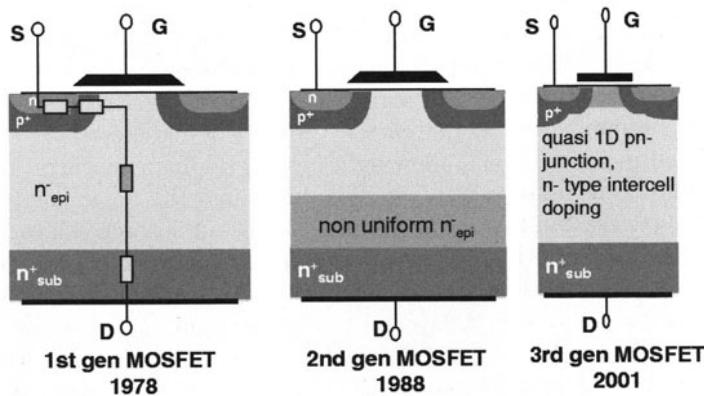


Fig. 16.3. Historical time line of conventional power MOSFETs

The manufacturers of power semiconductor devices produce both MOSFETs and IGBTs for SMPS applications. Nevertheless, as outlined above, both device concepts today have reached their respective physical barriers, which stand in the way of an optimization towards the ideal switch. The injected electron–hole plasma in the IGBT limits the reduction of switching

losses, whereas the MOSFET has its drawback in conduction losses due to its limited RD_{SON}.

Further optimization may therefore not be achieved by an evolutionary process, but it may be achieved by the adaptation of a new principle to power MOSFETs. This revolutionary solution is to use two doping profiles instead of one to optimize the performance of the device. The two carrier types are locally separated in the device: in the on-state the majority carriers, e.g. the electrons, yield a much lower RD_{SON} than in the conventional MOSFET, whereas in the blocking state the two carrier types cancel each other out, leaving a net doping level close to zero as required for a high blocking voltage. Owing to the fact that the charge required for low conductivity is not injected but built in, the device switches at least as fast as a MOSFET but with much better RD_{SON}*A. This device concept easily breaks the limit line of silicon, as the task of optimizing both the breakdown voltage and RD_{SON} with one doping profile no longer exists [5]. We call this principle, from to the compensation of p- and n-charges in the blocking state the compensation principle. Another name used is “superjunction device”, owing to the deep p-pillars which form a large pn junction [6]. This idea is derived from the field of lateral transistors, where it was originally developed back in the 1970s. The principle is known here as the RESURF (reduced surface field) concept [7]. Engineers from Philips observed in lateral transistors that the blocking voltage increased when was fabricated the structure within a very thin active n-layer on top of a p substrate or an insulating layer. The maximum electric field at the surface is then reduced owing to the relocation of the mirror charges away from the p-well towards the underlying p-layer. In other words, the charge associated with the n-doping of the drift region does not contribute to the field profile parallel to the current flow. The charge creates only a field component perpendicular to the current flow. Coe extended the idea later on to stacked p- and n-layers, allowing a higher doping concentration in the current-conducting n-layers [8]. Coe also described the basic compensation rules for the charge and the thickness dimensions of the stacked layers. As the stacked layers form another pn junction, the charge integral perpendicular to the interface must stay below the material-specific breakdown charge, which for silicon is $2 \times 10^{12} \text{ cm}^{-2}$. A layer of thickness 1 μm may therefore reach a doping level of up to $2 \times 10^{16} \text{ cm}^{-3}$, which is two orders of magnitude higher than the average doping level of a 600 V transistor. This easy calculation shows the potential of the charge compensation or RESURF principle. In lateral transistors, the RESURF principle is in widespread use today. Owing to the lateral current flow, however, the silicon volume is used to a relatively low extent. Lateral transistors are therefore used mainly in low-power applications such as chargers or standby power supplies. Medium- to high-power applications require a vertical transistor structure. The challenging task is therefore to adapt the idea of charge compensation to a vertical device concept and to develop a technology for manufacture. Even though the basic device structure

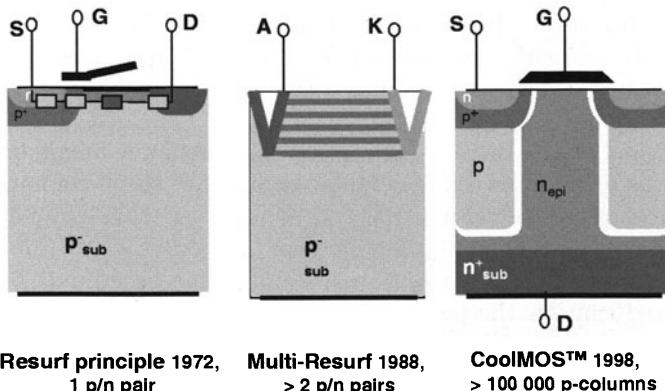


Fig. 16.4. Historical time line for the development of compensation devices

was discussed early on [9], it was never technically realized until the advent of CoolMOS™ in 1998 [5]. Figure 16.4 shows the basic device structures.

16.3 Manufacturing Technology and Its Challenges

The challenge lies in the formation of p-columns which reach deep into the drift region, e.g. to a depth of around 40 µm in a 600 V device. Today's commercial suppliers of compensation devices use a method based upon stacking n-doped epi layers over each other with the intermediate implantation of masked p-doped regions. These individual p-regions are finally merged into one p-column structure by diffusion. Figure 16.5 shows the process flow for this technique, called multiple epitaxy.

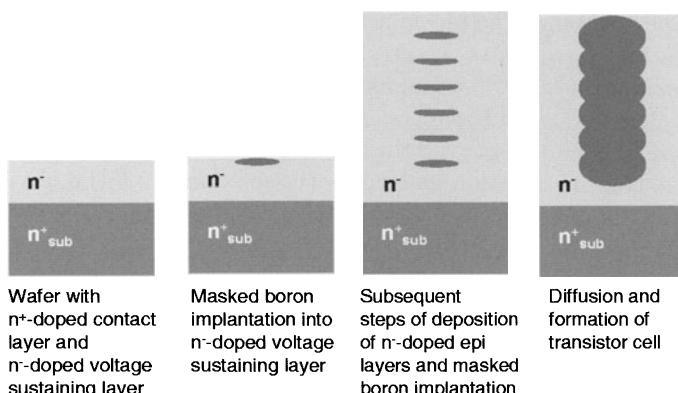


Fig. 16.5. Individual process steps required for the formation of the p-columns of compensation devices using the multiple-epitaxy approach

Besides this relatively expensive and time-consuming method, trench etching combined with epitaxial filling [10] or tilted sidewall implantations [11] has been suggested. Recently, ultra-high-energy ion beam projection has also been proposed [12].

As discussed above, there is a limitation caused by the breakdown charge, which acts as a link between the lateral spacing of the p-columns and the doping concentration of the n-regions in between. Let us assume as an example that we would like an n-doping level one order of magnitude better than the state of the art; our cell pitch would then be around 15 μm . This leaves up to 10 μm for the p-column and hence a maximum of 3 μm lateral outdiffusion per edge with an implantation window of 4 μm width. In this case we would also obtain 3 μm vertical outdiffusion in both directions, or in other words, an implantation region with a vertical depth of 7 μm . The formation of a 40 μm deep p-column would therefore require six such regions and hence six epi and implantation cycles in the manufacturing line. On the other hand, the n-conducting path would have 10 times the conductivity of a conventional MOSFET with the same blocking capability. In calculating the area-specific RDS_{on}*A we have to take into account the fact that the p-columns do not contribute to the forward conduction; we therefore obtain a 5 times improvement in RDS_{on}*A.

Another important aspect in relation to the manufacturing technology is the problem of charge balance [13]. Let us continue with the numbers of the above example. We start with an n-doping level 10 times the doping of a 600 V MOSFET with a pn^-n layer sequence. Our goal is to achieve at least the same blocking voltage with an unaltered drift region thickness. The absolute value of the net doping comprising the n-regions and the p-columns has therefore to stay within the limits of the original n^- doping; however, the net doping may be n-type or p-type. This is shown schematically in Fig. 16.6.

The blocking voltage is characterized by the area enclosed by the vertical-field distribution, whereas its slope is characterized, according to Poisson's law, by the net doping. The blocking voltage of a transistor with 10 times the doping of a 600 V transistor would be 100 V only (dotted line in Fig. 16.6); with the formation of the p-columns the net doping level decreases, and the blocking voltage rises and may reach 600 V (solid line in Fig. 16.6). With further increase of the p-charge the net doping level approaches zero, leading to a flat field distribution and hence the maximum blocking voltage. On the overcompensated side, the blocking capability starts to decrease, with the field distribution now being tilted towards the p-well instead of the n^+ contact (broken line in Fig. 16.6). Charge compensation devices therefore always show a strong relation between the blocking voltage and the absolute charge compensation. If we define a charge compensation degree κ by $(\text{n-doping} - \text{p-doping level}) / (\text{n-doping level})$ we obtain the desired breakdown voltage in the above example if the n-doping level is 10 times higher within a window of $\pm 10\%$. In other words, the gain G in RDS_{on}*A (which is roughly

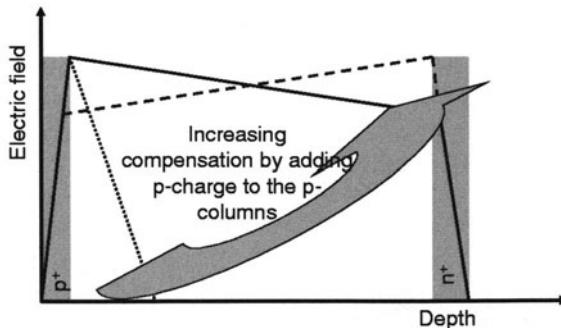


Fig. 16.6. Variation of the vertical electric field with increasing compensation of the n-columns by p-columns in an idealized compensation device. The p^+ and n^+ contact regions are shown in gray. The *solid line* shows the field distribution for the minimum charge balance (p undercompensating n), the *broken line* shows the field for the maximum charge balance (p overcompensating n) and the *dotted line* shows the field for no compensation with p-regions at all

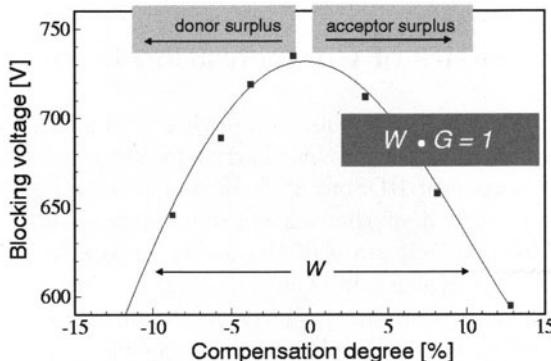


Fig. 16.7. The compensation parabola, showing the dependence of the blocking voltage on the compensation degree κ

half of the increase in the doping level) versus a conventional transistor, and the total process tolerance window W , in this case 0.2, are correlated by an easy formula $G * W = 1$. Figure 16.7 depicts the dependence of the blocking voltage on the compensation degree κ .

The potential of compensation devices for lower $RDS_{on} * A$ is therefore not limited by a physical boundary in the way that a conventional power MOSFET is limited by the silicon barrier, but depends only on the process control capability of the individual fab. This is the basis for the hope that continuous improvements in semiconductor technology will drive compensation devices along a roadmap with ever-decreasing $RDS_{on} * A$. Laboratory-scale prototypes show today the potential for 10 times improvement in $RDS_{on} * A$ at 600 V blocking voltage.

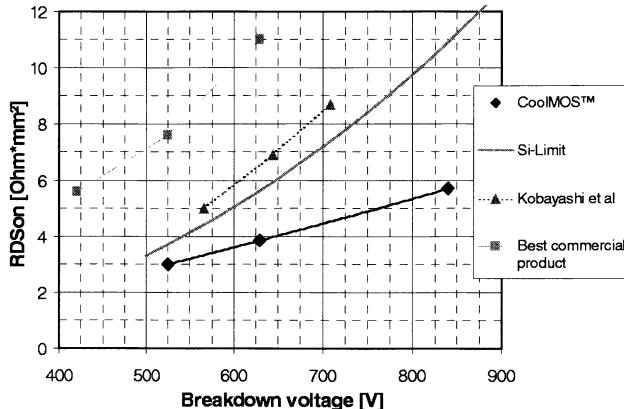


Fig. 16.8. The dependence of the specific on-state resistance on the blocking capability, showing the silicon limit for conventional MOSFETs, the best state of the art in commercial products and values in the literature in comparison with CoolMOS™ transistors

16.4 Characteristics of Compensation Devices

The main characteristic of compensation devices is the fact that they break through the silicon limit, the physical barrier for conventional power MOSFETs. The dependence of $RDSon^*A$ on breakdown voltage is defined in a nearly linear way instead of the conventional more-than-square law. Figure 16.8 shows today's best state of the art in conventional-technique and compensation devices. Kobayashi et al. [4] claim to come closest to the silicon limit; commercially available products from competitors are further from the silicon limit [14]. Recently released products based on the technology published by Kobayashi et al., however, hint at a maximum rating closer to the $RDSon^*A$ of its competitors than to the silicon limit [15]. These values are a clear indication of the barrier-like type of the silicon limit; an asymptotic approach is obtained ever-increasing efforts. A paradigm change requires a new structural approach. The points shown for CoolMOS™, the leading representative of compensation devices, are derived from the specific $RDSon$ maximum ratings at the minimum blocking voltages required for the 500 V, 600 V and 800 V product families, respectively.

This new law opens up new horizons, especially towards higher blocking voltages. Inside the device there is only one major difference from the conventional MOSFET, and that is the implementation of deep p-columns in the active zone or drift region of the transistor. During turn-on the current flows from the n^+ source region through the MOS channel, which may be adjusted laterally or vertically, and within the relatively high-doped n-region vertically towards the rear n^+ contact. In the on-state there is a small voltage drop across the drift region (in the order of some volts), which acts in

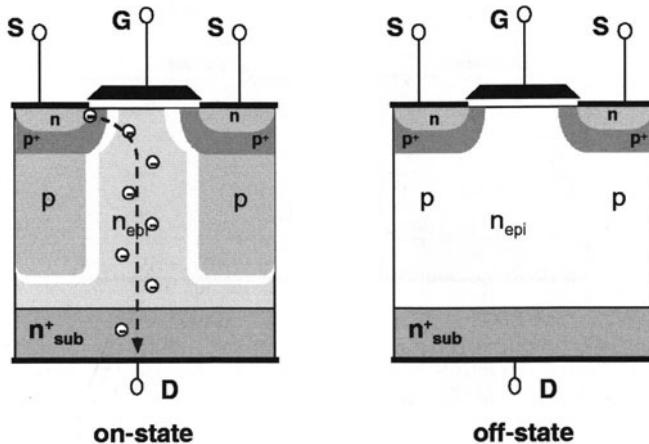


Fig. 16.9. Schematic representation of the on-state and blocking state of a compensation device

combination with the built-in potential as a reverse bias to the pn junction which is formed by the deep p-column in conjunction with the n drift region. We shall therefore always find a shallow depletion region around the deep p-column (see Fig. 16.9).

During turn-off the electron current is shut off by the MOS channel; the shallow depletion region around the deep p-columns starts to expand, and depletes the p-columns from the bottom to the top and the n-column from the top to the bottom.

This characteristic behavior is different from that of the conventional MOSFET, where the depletion region expands nearly one-dimensionally from top to bottom within the n-drift region. At a voltage of around 50 V – this value depends on the doping level and the compensation degree – the p- and n-column structure is totally depleted.

The device is now ready to take high blocking voltages. Note that in a conventional MOSFET we have at 50 V only a very small expansion of the drift region, which may cover less than 20% of the total drift region (see Fig. 16.10). The mobile carriers left in the drift region below the space charge region may lead to second breakdown effects triggered by a high slope dv/dt . The pileup of a high electric field in the depleted part of the drift region may lead to impact-ionization generation of more free carriers, which in turn may ignite the parasitic npn bipolar transistor that is always present in power MOSFETs. The early and complete removal of all mobile carriers from the drift region at a voltage of only around 10% of the total blocking capability makes compensation devices very rugged with respect to fast switching transients.

The compensation device combines two opposing device characteristics: on the one hand we have a very high conductivity, which is limited only by

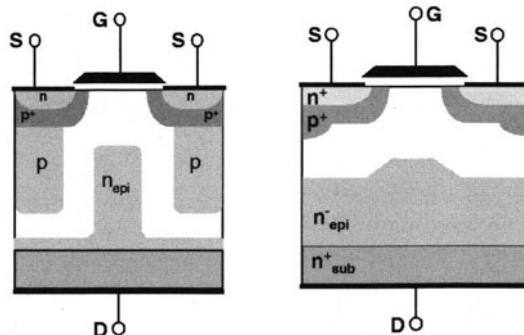


Fig. 16.10. Dynamic depletion of the drift zone during turn-off: comparison of a compensation device with a conventional MOSFET

process tolerances, as discussed, and on the other hand a blocking voltage with a fast turn-off characteristic like that of a MOSFET.

In the reverse blocking state the electric field distribution is characterized by three major field components. The vertical field is built up mainly between the p-well at the top and the rear n⁺ contact region. In the case of imperfect charge balance we obtain a modification of this vertical field component by uncompensated net charges in the drift region, which may be n⁻ or p⁻ type. This net charge leads to the compensation parabola shown in Fig. 16.7. The lateral pn junction formed by the deep p-columns with the n drift region in between superimposes a lateral field component on the vertical field distribution. Theoretical work has shown that in an optimum design the horizontal and vertical field components should be equal; in other words, the lateral charge integral across the column structure should not exceed 50% of the breakdown charge [6].

Let us return to the characteristic expansion of the depletion region in compensation devices: with increasing voltage we have, as in conventional MOSFETs, an increase in the space charge width, but we have additionally a decreasing surface area of the undepleted p- and n-columnns. As the p-columns are connected via p-wells to the source region, and the n-columnns are connected to drain potential, we obtain a strongly nonlinear drain-source, or output, capacitance of the device. Owing to the large three-dimensional p-column structure, the capacitance reaches a very large value at small voltages.

The same effect can be observed in the gate-drain, or feedback, capacitance. Here the leading effect is a quick expansion of the space charge layer down the n-column, which is mainly responsible for the strongly nonlinear behaviour. The gate-source, or input, capacitance can be reduced in comparison with conventional transistors owing to the potential for shrinking the area of compensation devices. Figure 16.11 shows such a comparison.

The energy stored in the output capacitance is usually transformed into heat during turn-on, if no measures are taken to ensure resonant switching

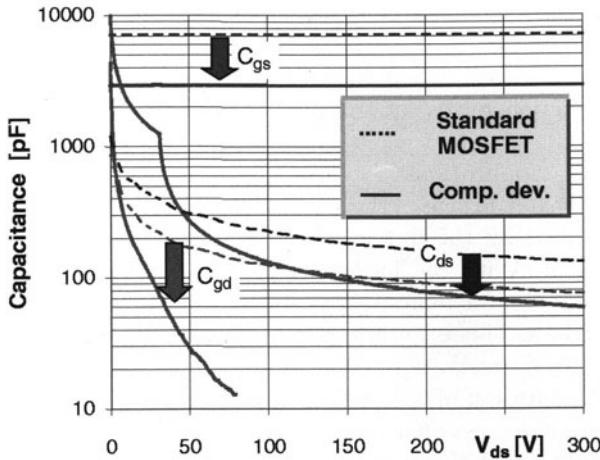


Fig. 16.11. Comparison of the device capacitances C_{gs} , C_{gd} and C_{ds} of a $190\text{ m}\Omega$ 600 V-rated compensation device (*solid lines*) versus a conventional power MOSFET (*dotted lines*)

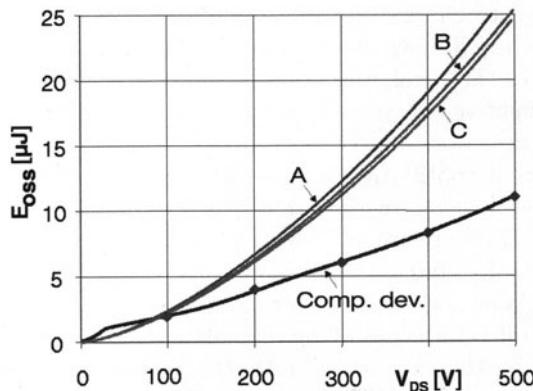


Fig. 16.12. Comparison of the energy E_{oss} stored in the output capacitance of a $190\text{ m}\Omega$ 600 V-rated compensation device (*diamond ticks*) with three conventional power MOSFETs (*solid lines A, B, C*)

transitions. This energy is therefore an important figure of merit for the calculation of switching losses. Figure 16.12 shows the corresponding figure, which is derived from the integral expression

$$E_{oss} = \int_0^{V_{DS}} C_{oss} \cdot U \cdot dU, \text{ where } C_{oss} = C_{ds} + C_{gd}.$$

The relatively low values of C_{gd} and C_{ds} at high voltage are therefore weighted with a strong emphasis relative to their large values at low voltage. At typical rail voltages of 350 to 420 V the compensation device shows a nearly 50% reduction versus conventional power MOSFETs from competitors (labeled A, B, C).

In resonant applications – especially the phase-shift zero-voltage-switching (ZVS) full bridge – the energy stored in the output capacitance has to be supplied by the flow of the load current through an additional inductive element. With the lower energy values of compensation devices, the bridge is capable of maintaining its ZVS characteristic down to lower load current values. This is an important design aspect in the application of these devices. The strongly nonlinear output capacitance, furthermore, helps during the commutation of the current from one leg to another [16].

Compensation devices offer unrivalled low on-state losses and low output energy. But what about the Joule losses during turn-on and turn-off? During turn-on the space charge, which is expanded across the whole p- and n-column structure, has to be flooded with mobile carriers. The electron charge is supplied via the load current flowing from the n^+ source region via the MOS channel. This current contribution is very fast and depends only on the time required to charge up the MOS gate. The p-charge, however, has to flow from the p-well down the p-column. This current contribution is also a drift current, as the p-column tends to float to negative potential during turn-on. This negative potential is derived from the negative charge of the uncompensated acceptors and is neutralized by holes flowing down from the p-well. This current contribution is also fast – within a few nanoseconds – if the compensation device provides a low-resistance ohmic current path down the p-column. It is therefore of crucial importance during manufacture to align the individual p-implantation islands in the proper way and to merge them sufficiently well together by thermal diffusion.

During turn-off both types of carriers are driven to their respective majority reservoirs by the expanding space charge region. No carriers have to cross the depletion region. In other words, the current and voltage are not in phase inside the device. The voltage rise is characterized by a soft curvature of the drain voltage below 70 V, which is due to the large, strongly voltage-dependent drain–source capacitance. As soon as the columns deplete each other completely the voltage rise becomes faster, and reaches values up to 80 V/ns without destruction.

Figure 16.13 shows simulated turn-on and turn-off characteristics for the model of the 500 V CoolMOS™ device used with TMA's MEDICI software. These simulated transients have the advantage of showing the real device performance without influences from the measurement setup or other external circuitry.

Like all MOSFETs, compensation devices have an internal body diode formed by the p-well and p-column on one side and the n-column and n^+

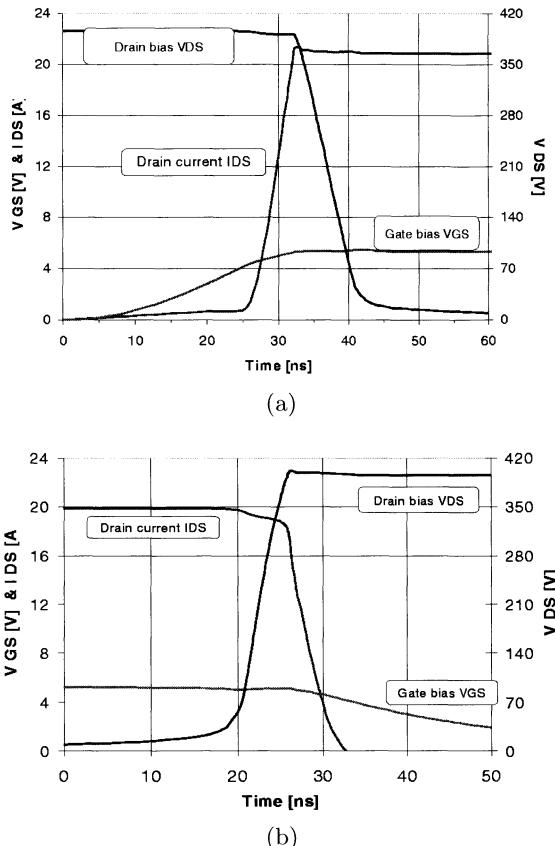


Fig. 16.13. (a) Simulated turn-on characteristics of a 500 V CoolMOS™ transistor cell; (b) simulated turn-off characteristics of a 500 V CoolMOS™ transistor cell

rear contact on the other side (see Fig. 16.9). The device is therefore capable of conducting current in the reverse direction. Owing to the deep p-column, the forward voltage drop of this body diode is only slightly above the threshold voltage of the junction and hence a good deal lower than in a conventional MOSFET with its large, low-doped n drift region. Of much greater importance, however, is the commutation behavior, which is the dynamic response of the device to the application of a reverse blocking voltage across the flooded body diode. The characteristic parameters here are the reverse recovery time t_{rr} ; the stored charge within the diode, usually called the reverse recovery charge, Q_{rr} ; the reverse recovery current peak I_{rr} ; and the softness factor S , which is a measure of the ratio of the charge extracted after and before the maximum reverse recovery peak (ideally 1). Unfortunately, compensation devices are not among the best in all of these parameters in comparison with

conventional devices. The optimization of compensation devices towards the lowest $RDS_{on} * A$ and a fast switching speed is contradictory to the optimization of the behavior of the body diode. Under forward bias the pn column structure is flooded completely by the electron–hole plasma. Owing to the large surface area of the pn junction and the relatively high doping level, we obtain a higher plasma concentration than in conventional MOSFETs and hence a larger Q_{rr} . When the device is reverse-biased, the plasma has to completely leave the active area before the pn column starts to reappear and starts to become depleted itself. As mentioned, the depletion commences within the p-column at the bottom, within the column structure along the side faces. The device is therefore not capable of sustaining a substantial blocking voltage before the plasma is nearly completely removed. This leads in turn to an untamed pileup of reverse recovery current and an extremely high reverse recovery current I_{rr} ; note that in a conventional MOSFET the device is capable of blocking a voltage at an early stage of commutation and therefore may slow down the rise in the reverse recovery current rise quite soon. When the compensation device starts to take the voltage nearly all mobile carriers have gone; this leads in turn to a very rapid discontinuation of the reverse recovery current and hence a softness factor S close to zero.

The diode performance may be improved by electron irradiation, with the reverse recovery time t_{rr} and charge Q_{rr} being reduced by factors of 3 and 10, respectively, but the softness and the overall behavior remain unaltered [17]. Figure 16.14 shows a comparison of the commutation of an irradiated and an unirradiated compensation device at a commutation speed of $100 \text{ A}/\mu\text{s}$, a speed that is usually used for datasheet ratings.

The diode is, however, very robust at high commutation speeds and has shown excellent performance in phase-shift ZVS bridge applications [18]. This again is due mainly to the absence of mobile carriers during the voltage rise and, additionally, to the separation of charges deep within the active region by the horizontal electrical field. The device therefore shows no tendency towards second breakdown effects or triggering of the parasitic bipolar transistor structure. Measurements have shown commutation speeds of up to $3000 \text{ A}/\mu\text{s}$. Figure 16.15 shows a comparison of the commutation behavior of unirradiated $600 \text{ m}\Omega$ transistors at a commutation speed of $1000 \text{ A}/\mu\text{s}$. The commutation is now completed within only 200 ns. The conventional device shows a large tail current after reaching the rail voltage of 350 V in this case. This clearly shows that the space charge region has not yet reached the rear n^+ contact. The remaining charge in the drift region may therefore trigger the parasitic npn transistor. The compensation device does not show any tail current; the charge is removed in an early phase of the commutation.

Note, furthermore, the nearly ideal softness factor S at this very high commutation speed. This phenomenon can be explained in a first-principles approach like this: the current discontinuation remains, independent of the commutation speed, more or less unaltered as it does not depend on the

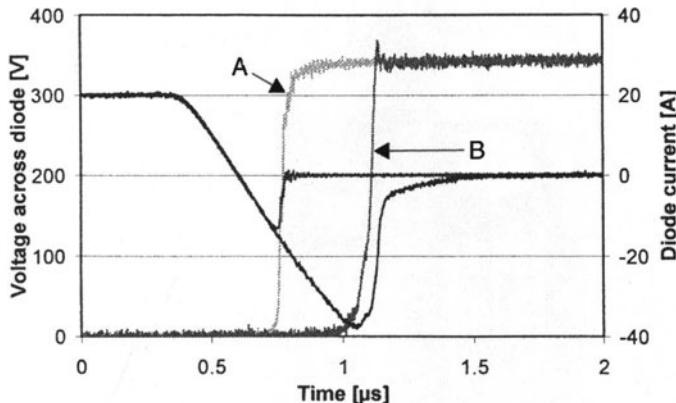


Fig. 16.14. Comparison of the commutation behavior A of an irradiated and B an unirradiated compensation device at a commutation speed of $100 \text{ A}/\mu\text{s}$ (600 V, $190 \text{ m}\Omega$ type) A, B drain bias

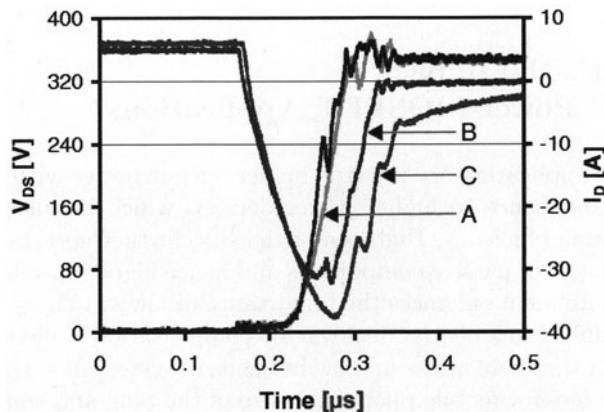


Fig. 16.15. Comparison of the commutation behavior of a compensation device versus a conventional transistor at a very high commutation speed of $1000 \text{ A}/\mu\text{s}$ (500 V, $600 \text{ m}\Omega$ types) A bias voltages V_{DS} , B compensation device and C conventional device currents

reverse recovery charge Q_{rr} but only on the device structure, unlike the situation for a conventional MOSFET. The performance before the reverse recovery current reaches its maximum, however, is a function of the commutation speed only. The softness factor S for compensation devices can therefore be greatly improved by switching faster and faster.

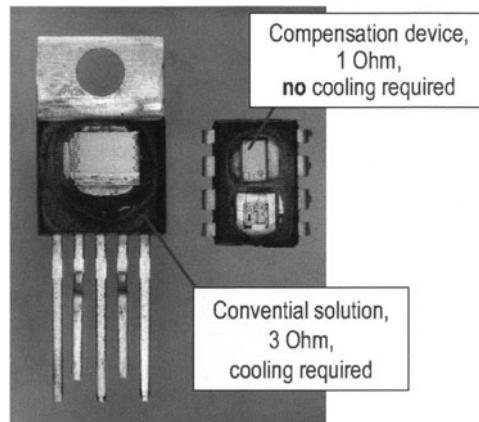


Fig. 16.16. Comparison of two system solutions: the conventional solution shows total losses of 3 W and therefore requires a heat sink; the solution based on a compensation device has total power losses of only 1 W and does not require cooling

16.5 What's the Impact on Typical Power MOSFET Applications?

The trends in application are towards higher output power with smaller system profiles and hence to higher power density, which in turn requires an improved system efficiency. High power density, furthermore, requires a reduction in the size of passive components and hence higher switching frequencies. This requirement enhances the importance of low switching losses. Both targets are fulfilled in a nearly ideal way by compensation devices. This novel device concept therefore opens up new horizons in system design and integration, e.g. chargers for mobile phones the size of the plug and small notebook adapters with an output power of 120 W or beyond. Furthermore, entire systems may be integrated on a package level. Figure 16.16 shows a comparison of a combined controller and power MOSFET made by conventional techniques with a similar combination made using compensation devices.

Finally, let us take a look at a typical example of an application, namely the power factor correction stage widely used today in SMPS and lighting applications. Figure 16.17 shows a typical SMPS topology, comprising a rectifier bridge, a power factor correction (PFC) stage and a pulse width modulation stage, here using a flyback converter topology. The advantage of using a PFC stage is that current is drawn continuously from the power line with a nearly ideal power factor of 1. The PFC stage may run in a continuous mode (CCM) or discontinuous mode (DCM), according to whether the current through the inductance discontinues between two switching cycles or not. The CCM mode requires a smaller peak current and may therefore use smaller MOSFETs but needs an extremely fast-reverse-recovery diode for the

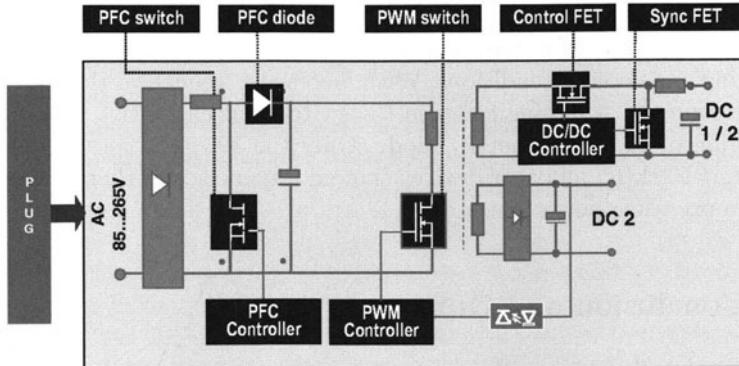


Fig. 16.17. Typical block diagram of a switch mode power supply for AC to DC power conversion, comprising line rectification, a primary-side power factor correction stage, a primary-side pulse width modulation stage and a secondary-side DC control stage

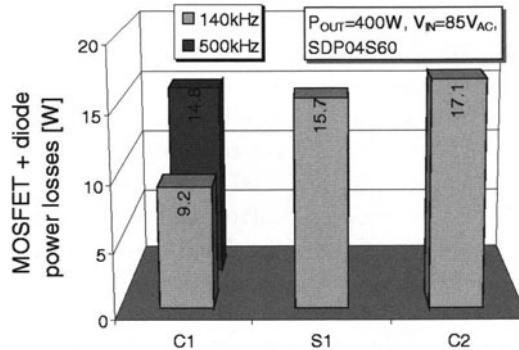


Fig. 16.18. Comparison of the total power losses of two compensation devices (C1, 600 V, 190 mΩ type; C2, 600 V, 380 mΩ type) versus a conventional power MOSFET (S1, 500 V, 230 mΩ) in a 400 W CCM PFC converter. For better identification of the performance of the power MOSFET, a SiC Schottky diode was used as a freewheeling diode in all tests

rectification. Combined with better EMI performance, the CCM mode is the right choice in high-power-density applications.

Recently, several studies have investigated the performance of Cool-MOSTM transistors in combination with ultrafast SiC Schottky diodes [19–21]. The absence of any reverse recovery charge in the SiC Schottky diode greatly reduces the losses in the associated power MOSFET and thus greatly enhances the efficiency of the PFC stage. Regarding the power MOSFET, fast switching and a low energy stored in the output capacitance are of primary importance. Compensation devices are excellent in both respects; they are therefore the first choice in hard-switching CCM PFC applications. Figure 16.18 shows a comparison of two different compensation devices (a

190 mΩ and a 380 mΩ type) versus a conventional power MOSFET (230 mΩ). All devices were tested in the same socket within the PFC board with identical driving and cooling conditions. With increasing frequency, the total losses increase owing to switching losses in the switch plus capacitive charging and discharging of the SiC Schottky diode. Only the 190 mΩ type is capable of running at 500 kHz; all other devices showed a tendency to thermal runaway owing to excessive power losses.

16.6 Conclusion and Outlook

Compensation devices are the most modern representatives of and the summit of the long development history of power MOSFETs. Owing to the novel application of the charge compensation principle, the long-existing relationship between RD_{SON}*A and the blocking voltage – the so-called silicon limit – is no longer valid. This opens up the possibility of revolutionary low-ohmic power MOSFETs in a given package outline or drastically smaller switches at a given RD_{SON}. Compensation devices show drastically reduced switching losses together with improved ruggedness towards high switching transients. Owing to these device characteristics, compensation devices are the ideal choice for advanced products in the area of power management and supply.

Is there another physical boundary to further reduction of the area-specific RD_{SON}? We have discussed the influence of charge compensation and hence the strict requirements on process control, which are of a severity hitherto unknown in the manufacture of power semiconductor devices. This is the main limitation today. But what about current density or premature depletion of the p- and n-columns?

A further reduction of the area-specific RD_{SON} requires a higher doping level of the current-conducting n-columns and – owing to the limitation imposed by the breakdown charge – a smaller lateral distance or pitch between adjacent n-columns or p-columns. If a new device family has to carry the same current as its predecessor with the same RD_{SON}, the current density increases at the same rate as the doping level of the n-columns. In this case we always maintain the same ratio between the charge associated with the current flow and the impurity concentration of the n-doping. In a good design this ratio should not exceed 10%, as the electric field should always be dominated by the impurity concentrations and not by mobile carriers. This so-called Kirk limit seems, therefore, not to be an obstacle on the way to further progress. So what about the early depletion of the p- and n-columns? With a reduction of the pitch and of the cell sizes, the depletion voltage will definitely go down. Today the columns deplete around 50 V, which is sufficiently distant from the voltage drop under typical on-state operation conditions of power MOSFETs. If the area-specific RD_{SON} is reduced by another order of magnitude relative to today's state of the art, this boundary becomes important. I would expect that in this case some application requirements will move towards lower current

levels at a given RDSon. These applications would then make use of the ultra-low RDSon that these future devices may offer, and hence lower the total power losses to an extent that cannot be predicted today. Other applications will stay with more conventional peak current requirements and will therefore use more conventional power MOSFETs, maybe compensation devices of the first generation. The market for standard power MOSFETs that do not make use of the compensation principle will definitely be small in 10 years' time. The race towards the best compensation device has already begun [22–24].

Acknowledgments

The author would like to express his gratitude to H. Weber, J.-P. Stengl, H. Strack, J. Tihanyi and the entire CoolMOS™ development team. Without their never-fatiguing efforts the compensation device technology would never have become a reality. We are also very proud to have been awarded the German Industry Innovation Award in 2001 for Infineon's proprietary IGBT and CoolMOS™ technology.

References

1. T. Laska, M. Münzer, F. Pfirsich, C. Schaeffer, T. Schmidt: The field stop IGBT (FS-IGBT) – a new power device concept with a great improvement potential. Proc. ISPSD (2000) pp. 355–358
2. T. Laska, M. Matschitsch, W. Scholz: Ultrathin-wafer technology for a new 600V-NPT-IGBT. Proc. ISPSD (1997) pp. 361–364
3. X.B. Chen, C. Hu: Optimum doping profile of power MOSFET's epitaxial layer. IEEE Trans. Electron Devices **ED-29**, 985 (1982)
4. T. Kobayashi et al.: High-voltage power MOSFETs reached almost to the Si limit. Proc. ISPSD (2001) pp. 99–102
5. G. Deboy, M. März, J.-P. Stengl, H. Strack, J. Tihanyi, H. Weber: A new generation of high voltage MOSFETs breaks the limit line of silicon. Tech. Digest IEDM (1998) pp. 683–685
6. T. Fujihira: Theory of semiconductor superjunction devices. Jpn. J. Appl. Phys. **36**, 6254 (1997)
7. A.W. Ludikhuize: A review of the RESURF technology. Proc. ISPSD (2000) pp. 11–18
8. David J. Coe: US patent 4,754,310 (1988)
9. J. Tihanyi: A qualitative study of the DC performance of SIPMOS transistors. Siemens Forsch.- u. Entwickl.-Ber. **9**, Nr. 4, (1980) pp. 181–189
10. M. Rüb, D. Ahlers, J. Baumgartl, G. Deboy, W. Friza, O. Häberlen, I. Steinigke: A novel trench concept for the fabrication of compensation devices. Proc. ISPSD (2003) pp. 203–206
11. T. Nitta, T. Minato, M. Yano, A. Uenishi, M. Harada, S. Hine: Experimental results and simulation analysis of 250 V super trench MOSFET. Proc. ISPSD (2000) pp. 77–80

12. J. Meijer, B. Burchard, K. Ivanova, B.E. Volland, I.W. Rangelow, M. Rüb, G. Deboy: High energy ion projection for deep ion implantation as a low cost, high throughput alternative for subsequent epitaxy processes. Proc. EIPBN (2003)
13. P.M. Shenoy, G. Dolny: Analysis of the effect of charge imbalance on the static and dynamic characteristics of the super junction MOSFET. Proc. ISPSD (1999) pp. 99–102
14. ST Supermesh product family, values calculated from best-of-class type 600 V, 550Ω max. rating
15. Fuji FAP-G product family, values calculated from best-of-class type 600 V, 540Ω max. rating
16. M. Schutten: General Electric, Schenectady, personal discussion
17. M. Schmitt, H.-J. Schulze, A. Schlägl, M. Vossebürger, A. Willmeroth, G. Deboy, G. Wachutka: A comparison of electron, proton and helium ion irradiation for the optimization of the CoolMOSTM body diode. Proc. ISPSD (2002) pp. 229–232
18. G. Deboy, J. Hancock, M. Pürschel, U. Wahl, A. Willmeroth: Compensation devices solve failure mode of the phase shift ZVS bridge during light load operation. Proc. APEC (2002)
19. L. Lorenz, G. Deboy, I. Zverev: Matched pair of CoolMOSTM transistor with SiC Schottky diode – advantages in the application. IEEE IAS Conf. Rec. (2000) pp. 376–383
20. I. Zverev: Frequency related trade off's in a hard switching CCM PFC boost converter. Proc. APEC (2003)
21. B. Lu, W. Dong, Q. Zhao, F.C. Lee: Performance evaluation of CoolMOSTM and SiC diode for single-phase
22. M. Saggio, D. Fagone, S. Musumeci: MdmeshTM – innovative technology for high voltage power MOSFETs. Proc. ISPSD (2000) pp. 65–68
23. Y. Onishi, S. Iwamoto, T. Sato, T. Nagaoka, K. Ueno, T. Fujihira: 24 mOhmcm^2 680 V silicon superjunction MOSFET. Proc. ISPSD (2002) pp. 241–244
24. W. Saito, I. Omura, S. Aida, S. Koduki, M. Izumisawa, T. Ogura: 600 V Semi-superjunction MOSFET. Proc. ISPSD (2003) pp. 45–48

17 Integrated Circuits

J. Borel

17.1 Introduction

Integrated circuits (ICs) started to become a reality in the early 1960s with the advent of small-scale integration of bipolar transistors NAND gates. The baton was quickly passed the CMOS technology, which provided simpler processes, lower costs and greater expectations in terms of integration capabilities. The time spent on and the cost of design were rather low owing to the low complexity of the functions, a situation that has completely changed today.

The applications of silicon need to be understood in terms of market demand, but their market success is so dependent on other parameters that we have to describe those parameters before addressing the products and their market.

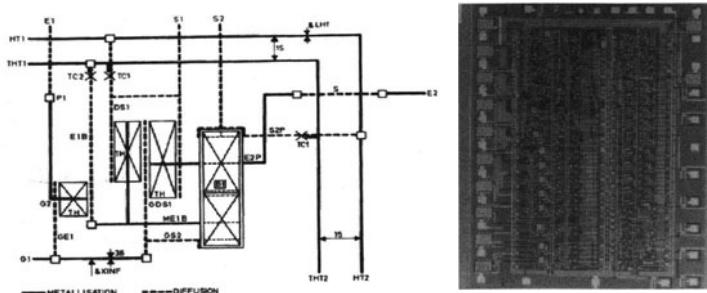
17.2 A Jump into the Past

A key transition period in the history of silicon was the years 1960 to 1970, when small-scale bipolar products (10 to 20 gates) were on the market and the CMOS, complementary MOS process was in its early design phase. The engineers were facing a difficult challenge in their designs: to move from devices with physics-defined parameters (the input diode of the bipolar transistor) to devices with technology-dependent parameters (the drifting oxide threshold of the CMOS device). Design methodologies using CMOS devices were felt easier to implement owing to simpler layout constraints than for bipolar devices.

An example of a product with its layout methodology is given Fig. 17.1 [1]. Following the start of production the process improved steadily, and a new product era started with the challenge, from the beginning, of making more complex products at lower cost with this new technology.

Since those early times, this driving force for CMOS has remained the same. The result has been an incredible shrinking of critical dimensions of the devices versus time (as forecasted in the ITRS, international technology roadmap for semiconductors), together with an increase in the area of silicon wafers, as shown in Fig. 17.2.

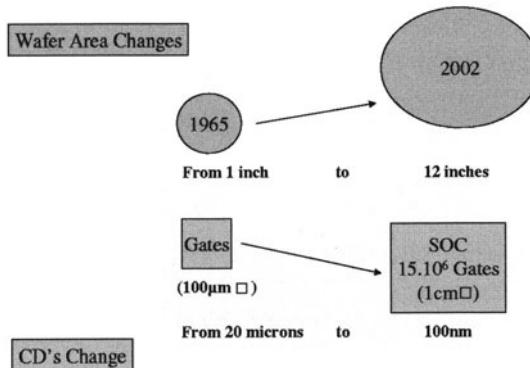
Automatic Layout Synthesis Example (l'Espalier 1975)



Date 2002 : Paris, March 4-8

15

Fig. 17.1. Example of a product and its layout methodology



Date 2002 : Paris, March 4-8

6

Fig. 17.2. Shrinking of critical dimensions and increase of wafer area

The expectations in terms of a decrease in final product cost have been met well beyond the early forecasts. As a rule of thumb, for memories, we can say that the cost of a square millimeter of processed silicon has been kept nearly constant over time and that the cost decrease per bit is equal to the corresponding density increase factor (six decades in 30 years).

17.3 Importance Gained by ICs in the Global Economy

Integrated circuits have been the fuel of the global economy in the sense that their pervasiveness has been seen successively in all their sectors of application (Fig. 17.3):

New Market driving forces

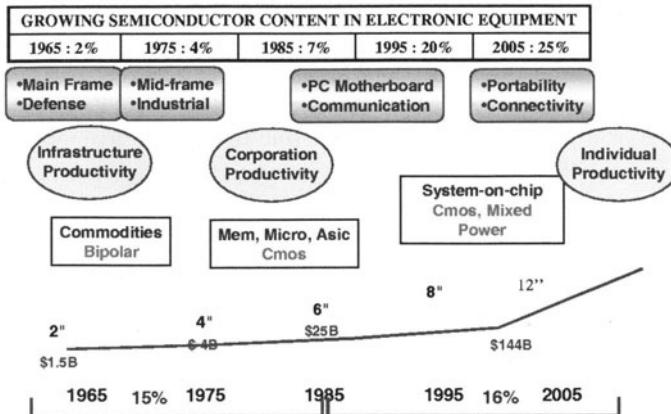


Fig. 17.3. The growing pervasiveness of ICs, integrated circuits. Source, ST

- In the 1960s their first contribution was mostly in the commodity market (mainframes and mid-frame computers) for the defense and industrial sectors; here, bipolar technologies were used.
- The move to CMOS in the 1980s gave higher integration capabilities and the availability of memories and cores (microprocessors), which were used in an ASIC (application-specific integrated circuit) design methodology. Standard cells and seas of gates started to be used to speed up the whole process of getting to the final product (higher density was achieved with ASICs in standard cells, whereas there was faster prototype availability with the sea of gates but at higher real-estate cost).
- Driven by process enhancements (shrinking dimensions and using multi-layers of interconnections), the move to what is called SoC (system on chip) [2] started to become fully operational at the beginning of the new millennium (the year 2000).

The reason for that move was mostly to address questions of individual comfort and productivity (being connected to the rest of the world any time and anywhere, at will, at an affordable cost to the user). This trend is continuing and will make additional progress in its applications thanks to new breakthroughs in technology, namely low-power, integrated radio frequency (RF) devices and microelectromechanical (MEMs) devices.

This technology will aim at giving full connectivity, assistance, safety and security to every single individual, at low cost, and will deeply influence our way of living.

17.4 The Market Constraints

Low cost per function has been since the beginning, and remains, the main driver for technology evolution, thanks to collective (i.e. several devices and processes are realized on the same wafer) manufacturing. Other benefits of this trend are reliability, lower weight in the case of mobile assistants, and higher speed or lower power consumption due to smaller geometries and consequently lower parasitic capacitances.

One typical example of this cost decrease is well known in the DRAM, dynamic random access memories market, where every eighteen months the cell size is decreased by a factor of two and a corresponding scaling is seen in the decrease of the price per bit; the cost of processed silicon per square millimeter has been nearly constant over the years explaining the drastic decrease in price per bit.

The other applications (digital, analog and mixed) have benefited from such an evolution of the memory process but are subject to additional specific constraints arising from the higher complexity of their technology, their design and their testing.

These are the roots of the present “digital revolution”, where nearly everything that can be handled, transmitted or processed in digital form. This leads to systems that give the best trade-off in performance and cost in our analog world (to which such systems need to be interfaced).

One aspect of the cost to the user can be seen through the “time to market” (TTM) parameter, which is a key differentiator in competitiveness for semiconductor suppliers [2]. This situation is represented in Fig. 17.4, where two competitors A and B are addressing the same market:

- One is in phase with the market (curve A).
- The other has a delay D in its entry into the market (curve B).

Time to Market Revenue Penalties

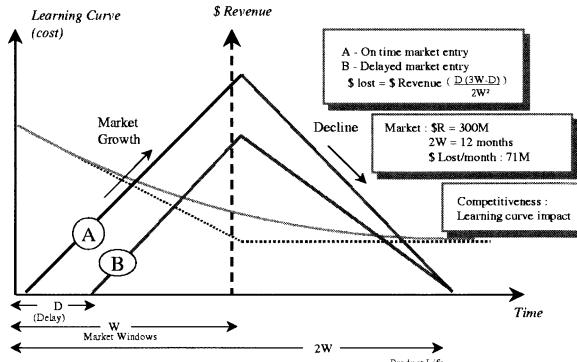


Fig. 17.4. Effect of time to market on competitiveness

Also represented is the learning curve, that is to say the additional advantage of supplier A due to the decrease in manufacturing costs for its products (as a rule of thumb, production costs are divided by two each time the manufactured quantities are multiplied by ten).

The curve clearly shows the market dominance that can be acquired using the short-TTM approach, even if higher R&D investments are usually necessary in this case:

- The serviceable available market (SAM) is larger.
- The gross margin is higher because of the decrease of manufacturing costs.

This leads to a market dominance for companies able to achieve a short TTM (e.g. Intel in the case of microprocessors).

17.5 The Products “Enablers”

There are a few “enablers” that are necessary to create competitive products (an “enabler” is defined here as a technique, i.e. a process or design automation, necessary to create a product):

- The semiconductor process: it should be ahead of the competition to provide the best products (versus performance) at a lower price (high density).
- Design automation is becoming a key factor in the competitiveness of semiconductor companies, as shown in a recent IBS study [3]. This has long been argued but is now clearly understood with the advent of 90 nm technologies, where even the process itself will not be engineered without the availability of powerful EDA, electronic design automation tools (for defect observability, design centering, design for manufacturability, design for robustness, . . .).
- IPs (intellectual property) libraries are necessary in a platform-based design approach [4], where we try to reuse as much as possible the functions already designed in a previous process. In that domain, we are still far from agreed standards, though some significant advances have been made recently [5].

Such an evolution allows one to put more and more functionalities on a single chip, and the addition of mixed analog–digital functions gives access to the SoC. Even though there is an increase in the complexity of the process and the design, for large-volume applications this approach gives rise to consumer products never achieved before (Digital Versatile Disk, Walkman, video cameras, digital photography, electronic agenda . . .). Even MEMs are starting to appear in SoC chips to improve the performance (switches, inductances, sensors, actuators, . . .).

An example of an SoC product for multimedia is shown in Fig. 17.5 with some of its global functions.

System on Chip Challenge

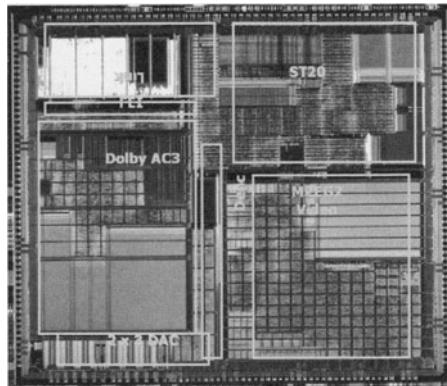


Fig. 17.5. Example of an SoC

Last but not least, the major bottleneck that is seen in the next generation of products is in its specification itself, which is becoming very complex and will need to be formalized and validated in a high-level specification language. This language must allow one, before finalization of the specifications, to study “the usages” with the final customer on a virtual prototype. This will provide a formal system specification synthesizable within the design flow.

This approach becomes mandatory with 90 nm designs for reasons of time to market and also of cost (the cost of one mask set exceeds \$1 million)

17.6 Integration Capabilities

In the short term, technology integration capabilities will continue to increase to include multiprocessor systems, large amounts of memory and mixed signal functionalities to interface with the analog world. The achievement of performances at reasonable power consumption will drive new applications or allow integration of more and more functionalities.

Some evaluations of the performance of several generations of technology are given in the Table 17.1.

In 2004, with the 90 nm technology, an example of an application in the multimedia market that integrates on a single chip a home gateway and a set top box will contain the following functions:

- network interface (XDSL, undefined (digital or analog) digital subscriber line, cable, PON, passive optical network, ...).
- satellite television reception (DVB), digital video broadcast.

An evaluation of the chip content is given in Table 17.2 [6].

Memories will become the components most used on the chip (70% of the area), power consumption will be a sensitive parameter to be minimized and

Table 17.1. Technology capabilities (estimated)

Year	2000	2004	2011
Technology node	0.18 μm	0.09 μm	0.05 μm
Characteristics (for digital blocks)			
Gates/ cm^2	5×10^6	30×10^6	200×10^6
GOPS/ cm^2	70	370	2550
MIPS/watt	1000	2200	4000
SRAM Mbit/ cm^2	60	120	240
FMAX ^x	< 720 MHz	< 1300 MHz	< 2 GHz

^x $\text{FMAX} = 1/18 \tau_p$ (optimal silicon performance)

GOPS, giga operations per second

MIPS, million of instructions per second

SRAM, static random access memory

FMAX, maximum frequency

τ_p , propagating time per gate

Table 17.2. Home gateway example (estimated)

Technology node	Analog	logic	IP reuse		Memory	Comments
			DSP + Copro	MICROP		
Y2004	Area	1 mm^2	3 mm^2	15 mm^2	71 mm^2	100 mm^2
	No. of gates	–	10^6	2.5×10^6	1.7×10^6	–
	GOPS	–	–	55 GOPS	–	–
	GIPS	–	–	–	37 GIPS	•
	Mbits	–	–	–	–	85 MEG
	RAM	–	–	–	–	8 SRAM + 64 NV
	Watts	\ll	1 [#]	5.1 [*]	3.7 [*]	+ 245 DRAM

Average: # -25% of active logic (partial power down)

* -10% use of processors (at Fmax)

• Equivalent complexity to pure SRAM

GIPS, giga instructions per second

DSP, digital signal processor

MICROP, Microprocessor

NV, non volatile memory

multicore integration will be common (some commodity products already have more than 10 processors).

What is not shown in this table is the huge development cost of the underlying embedded software needed to develop such complex applications. In this domain, a lot of progress remains to be made in increasing programming efficiency.

17.7 Design Bottlenecks

Design integration capabilities in 2011 will reach 200 million gates (or SRAM bits) per square centimeter. Even if the product specifications are formally defined in a given design, huge quantities of data will need to be handled, validation at every level will be needed (today this requires 70% of the design time) and the design must be good at the first prototype to cope with market goals. A summary of the conflicting constraints in understanding the global picture and mastering the details is given in Fig. 17.6.

Among the priorities in establishing an SoC design flow are the following:

- standardization of high-level design languages (for usage validation and formal specifications);
- creation of a full top-down system design flow with constraint propagation (see [6]) and interconnect-centric design;
- development of libraries of IPs to reuse the already integrated functionalities as much as possible;
- improvement of software design efficiency, the fastest-growing resource-demanding task today;
- development of a full design environment for signal integrity, a new bottleneck in design integration.

Semiconductor companies will have to invest heavily in new design methodologies and design automation tools. As mentioned earlier (see [3]), the successful companies will be those which anticipate their needs for implementation of new design solutions and for training of their designers (up to 5 years ahead of the actual needs).

The design teams that face the complexities of the tasks and of the tools are, more and more, organized into local competence groups to address large projects. Their main domains of competence are:

- system architecture and hardware/software codesign;
- verification;

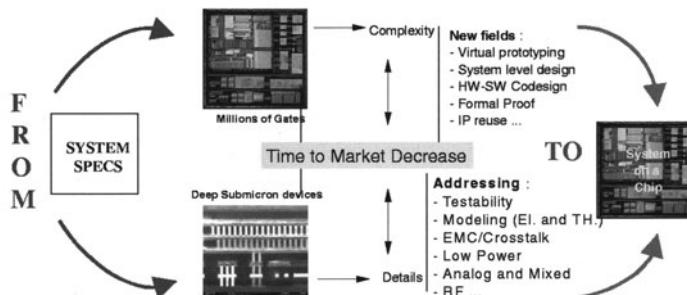
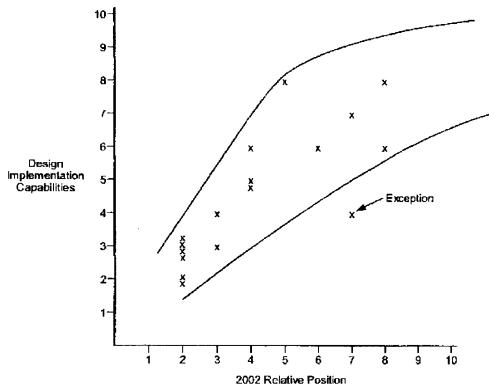


Fig. 17.6. Constraints in design integration

ASIC COMPANIES

Relationship Between 1998 Customer Rankings in Design Implementation Capabilities and 2002 Relative Position



Source IBS

Medea+ Roadmap Meeting
January 16th 2003

4

Fig. 17.7. Correlation of performance with investment in design capabilities

- IP creation;
- mixed A/D and RF design;
- signal integrity;
- low power.

The results of the above-mentioned IBS study show a clear correlation of the performance of IDM, independent device manufacturers and ASIC, application specific IC, companies (both in performance ranking and market valuation) with their past investments in EDA (see Fig. 17.7) (a situation that will be even more critical in the coming years).

This figure shows a good correlation between the design implementation capabilities of companies in 1998 (the product of the percentage of revenue invested by an ASIC company in EDA with a factor from 0 to 10 representing how its customers value their EDA service organization) and their relative market position in 2002. It is therefore of great importance to anticipate needs in EDA to stay ahead of the competition, a rule that is still not clearly understood at high management levels.

17.8 Application Domains and Product Families

17.8.1 Application Domains

The main application domains in 2002 are briefly described below (see Fig. 17.8):

- data processing: includes desktops, servers and notebooks, and has the largest market size, with a CAGR (constant average growth rate) close to 10%;
- communications: includes cellular base stations and switching, cellular handsets, and other communications devices; it is driven by the Internet and portability;
- consumer: includes digital audio/video with DVD, digital video desk or digital versatile desk, digital set top boxes, color TV, VCRs, camcorders and digital still cameras;
- automotive: includes body system, power train and chassis;
- industrial: covers mostly power management;
- military and civil aerospace: remains low.

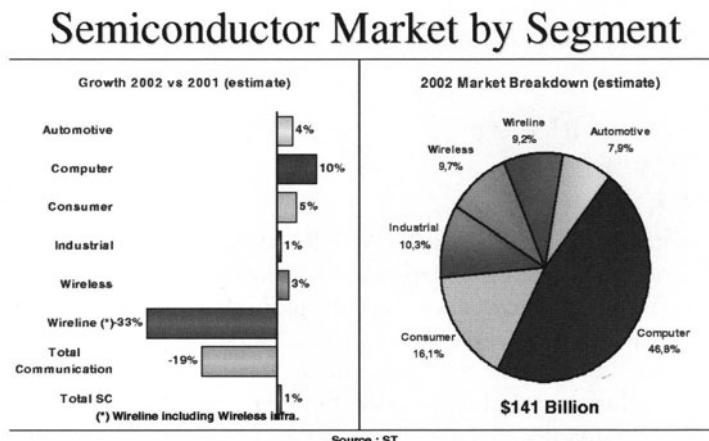


Fig. 17.8. The main application domains of semiconductors

17.8.2 Changes in Product Families

The semiconductor market covers four main categories of products:

- discrete, standards and optoelectronics;
- memories;
- microprocessors units (MPUs) and peripherals;

Semiconductor Market by Product

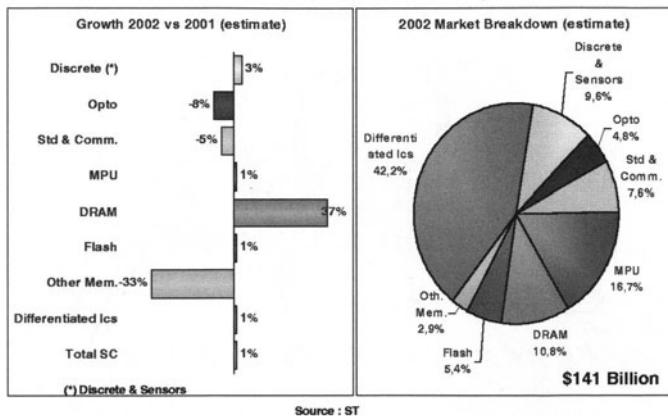


Fig. 17.9. Semiconductor market families

- differentiated ICs (application-specific standard products, or ASSP, application specific standard products and application-specific ICs, or ASICs) and microcomputer units (MCUs).

We are interested mostly in integrated circuits, which cover around 75% of the semiconductor market.

With the development of SoC, covering a significant part of the differentiated ICs, there will be a corresponding change in the market families (see Fig. 17.9):

- more and more memories will be embedded in the chip;
- more and more processors will be embedded in the chip;
- analog functions will be at the periphery of the digital chip.

We can foresee a significant potential for an increase of the SoC market in the coming years with a corresponding leveling of the other, traditional components of the market.

17.9 Conclusion

We are now familiar with the up and down cycles of the semiconductor business, triggered by overcapacity of the semiconductor production plants. Upturns and downturns will still occur owing to strong competition but they are likely to decrease in amplitude, with a higher contribution of foundries in production worldwide.

The potential bottleneck in the system will be in the capacity to generate new products in due time and such that they will be accepted by the market. This will require more design engineers, with much better design automation tools.

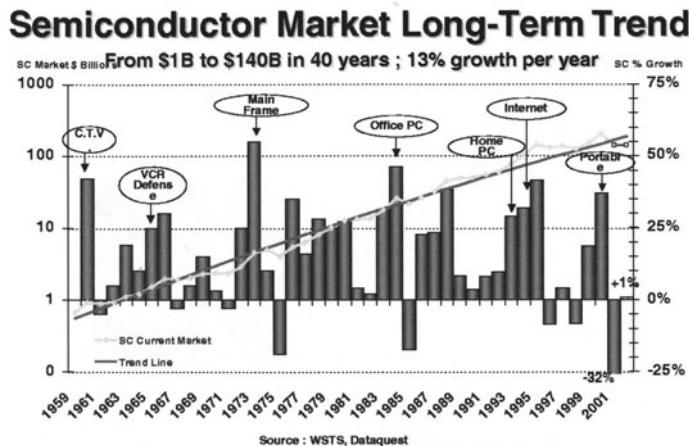


Fig. 17.10. Long-term trends in the semiconductor market

Beyond the cyclic aspect of the market mentioned earlier, we must not forget the megatrends that have brought this industry from \$1 billion in the 1960s to \$140 billion at the beginning of the new millennium, as shown Fig. 17.10.

It is interesting to note in this figure that some very large individual markets have been the drivers of the progress, and this should put even more emphasis on achieving a better understanding of the new markets by developing the “usages” with the future consumers before creating the new products.

References

1. Proceedings of Date 2002, Paris, March 4–8 (2002) at www.date-conference.com
2. J. Borel: System on silicon, where are we?, Keynote speech. Proceedings of the 14th VLSI Test Symposium (1996)
3. Analysis of the relationship between EDA expenditures and competitive positioning of IC vendors (International Business Strategies, Inc., 632 Industrial Way, Los Gatos, CA 95030, USA)
4. J. Borel: Technologies for multimedia systems on a chip, Keynote speech. Proceedings of the IEEE Solid State Circuits Conference (1997) pp. 18–21
5. IP, MEDEA+ application program at www.medeaplus.org
6. The MEDEA Design Automation Roadmap, 3rd release (2002), Design Automation Solutions for Europe, at www.medeaplus.org

18 Silicon Nanoelectronics: the Next 20 Years

L. Risch

18.1 Introduction

According to Moore's law, which predicts a decrease in feature sizes by a factor of 0.7 every 3 years, silicon transistors have become smaller and smaller in order to achieve higher integration densities, higher speed, lower power consumption and lower costs. This has been accomplished very successfully in the last 20 years, but will it continue in future? The latest ITRS, international technology roadmap for semiconductors (2001), describes in detail structural and electrical values for the scaling of CMOS down to the 22 nm node. Without any doubt, CMOS is considered to remain the mainstream technology for logic and memory. Many challenges have to be addressed for these CMOS generations regarding lithography, metallization, power dissipation and circuit design. Focusing on the device, the mandatory improvement in performance will be the key issue for further downscaling. Limitations arising from basic physical laws are still far away and will become important only well below sizes of ten nanometers.

Therefore, novel architectures for MOSFETs are needed, to improve the electrical parameters further and thus pave the way to much smaller transistors than those expected in the past. The most promising concepts beyond the conventional bulk Si MOSFET are vertical transistors and SOI, silicon-on-insulator devices, especially with a fully depleted channel and an ultrathin silicon substrate. Further improvements down to 10 nm and beyond will very likely be achievable by utilizing two or more gates for the control of the charge carriers. Moreover, mobility can be improved with SiGe and strained silicon materials. Realization of these concepts is much more likely than a replacement of CMOS by completely new devices such as single-electron transistors or molecular or carbon nanotube devices. Therefore, it is predicted that Si will achieve the 22 nm CMOS generation with gate lengths down to 9 nm in the year 2016 and will continue even further.

18.2 CMOS Scaling

The scaling of CMOS technology has made possible the key applications of electronics in our daily life such as the PC, mobile communication, the microprocessor, the Internet, and the various controllers in our cars and houses.

As we approach the sub-100 nm region many roadblocks to further reduction of feature sizes seem to appear, owing to limitations of lithography, lack of transistor performance, increasing interconnect delays and problems of power dissipation. Therefore, a slowdown in the evolution of semiconductor technology to smaller dimensions could be expected in the future. But in fact the opposite seems to be happening. In the last few years, an acceleration in the shrinking has been observed. This is because higher switching speeds have been achieved and the chip costs can be reduced owing to smaller die sizes. Whereas today the 90 nm CMOS technology has been ramped up to mass production in the most advanced semiconductor lines, a 22 nm CMOS generation is expected in the year 2016 according to the ITRS 2001 Roadmap [1], with gate lengths down to 9 nm for high performance, corresponding to 40 Si atoms in the 100 plane.

Looking at the devices, we see that today, except for some niche applications, bulk silicon is used with doped channels to obtain different threshold voltages of the transistors. Typically, n^+ / p^+ poly-silicon is the gate material, with silicidation, and SiO_2 is the gate dielectric. A spacer is deposited at the gate, with shallow source and drain extensions below to reduce short-channel effects. Finally, the contacts are provided by highly implanted As or B regions for n- or p-channels, respectively.

An SEM cross section of such a Si transistor with a gate length of 100 nm is shown in Fig. 18.1. Here silicon-on-insulator material with a Si thickness of 45 nm was used, but the processing and electrical characteristics are similar to those of bulk devices owing to the relatively thick Si layer.

Basically, a MOSFET acts like a switch. The off current should be in the range of pA to nA per micrometer of channel width and the on current should reach about 1 mA/ μm . Moreover, in basic CMOS circuits n- and p-channel devices are mostly connected in series, so the on current flows only during switching and therefore the standby power is low. Also very important is the ability of CMOS to always provide the full signal levels, i.e. 0 V or the power supply voltage.

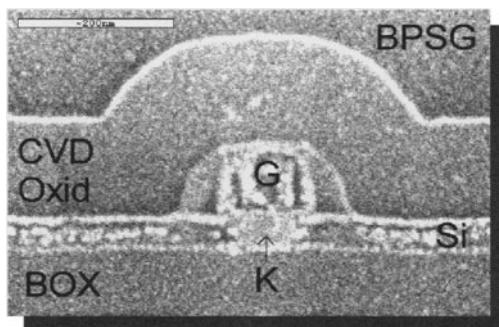


Fig. 18.1. Experimental 100 nm n-MOSFET on SOI with $t_{ox} = 2.5$ nm

year	01	04	07	10	13	16	19	22	25
node nm	130	90	65	45	32	22	18	13	10
L _{gp} nm	65	37	25	18	13	9	6	4	3
<i>t</i> _{ox} nm	1.5	1.2	0.9	0.7	0.5	0.4			
<i>x_j</i> nm	35	20	15	10	7	5			
N cm ⁻³ 10 ¹⁹	0.4	1.1	2.3	5.0	13	50	?CMOS ?		
V _{dd} V	1.2	1.0	0.7	0.6	0.5	0.4			
I _{off} μA/μm	0.01	0.1	1	3	7	10			
I _{on} mA/μm	0.9	0.9	0.9	1.2	1.5	1.5			
T _{dn} (C _V /I) ps	1.6	1	0.7	0.4	0.22	0.15			

Fig. 18.2. ITRS 01 device parameter requirements for high performance

Regarding further scaling of MOSFETs, many serious device challenges such as gate oxide thickness and junction depth are indicated by red brick walls in the ITRS Roadmap [1]. For example, for the 32 nm generation, the oxide thickness of the transistors should be only 0.5 nm, the junction depth 7 nm, the doping concentration $13 \times 10^{19} \text{ cm}^{-3}$ and the supply voltage as low as 0.5 V; see Fig. 18.2.

All these values seem difficult to achieve, but should not be regarded as physical limitations yet. In fact, this statement is more a serious suggestion to manufacturers that they should put enhanced effort into research and development on these issues. Otherwise, no performance improvement can be achieved with further scaling, owing to degradation of the on and off currents, leading to a slowdown of the intrinsic switching speed T_{dn} in the end. This value is extraordinarily fast and is in the range of picoseconds; it is given in simplified form by the gate capacitance times the power supply voltage divided by the on current.

The fundamental limits [2, 3] arising from quantum mechanical effects, for example, are still far ahead. One limit is given by the wave nature of the electron, which has a wavelength of about 5 nm in Si. Another limit is the energy quantization, which is a few meV at a size of 10 nm, and finally there is the atomistic limit of miniaturization, which is the spacing of the Si atoms in the crystal lattice, about 0.3 nm for the 100 surface. All these effects are still small and not really limiting for the next ten to twenty years. Consequently, in the future new Si MOSFETs with better performance than the conventional bulk transistors will have to be developed.

18.3 Novel MOSFETs Below 50 nm

18.3.1 Strained SiGe

The key device parameters for MOSFETs are low off and high on currents. In order to achieve the desired high on currents at decreasing power supply

voltages (see the ITRS Roadmap, Fig. 18.2), the improvement of the mobility of the charge carriers with SiGe heterostructures is becoming a very important issue. In silicon, the hole mobility is much lower than the electron mobility. Typical low-field mobilities measured in today's transistors with relatively high doping concentrations and thin gate dielectrics are in the range of $300 \text{ cm}^2/\text{Vs}$ and $100 \text{ cm}^2/\text{Vs}$ for electrons and holes, respectively. The first step towards high-mobility channels will be to improve the hole mobility in the p-channel transistor. This can be achieved by growing by epitaxy a thin $\text{Si}_{1-x}\text{Ge}_x$ layer, where x is the Ge concentration, with a thickness of 5 to 10 nm for the channel region, as shown in Fig. 18.3. For the p-channel transistor [4], the SiGe layer can be grown directly on bulk silicon using low-pressure CVD epitaxy at a temperature of 700 to 800°C. On top of the SiGe layer, a thin Si cap layer with a thickness of 3 to 5 nm is grown additionally. This forms a quantum well for the holes owing to the Si/SiGe/Si heterostructure, which is created by a step in the valence band with a depth of 150 mV for a Ge content of 0.2 (see Fig. 18.4).

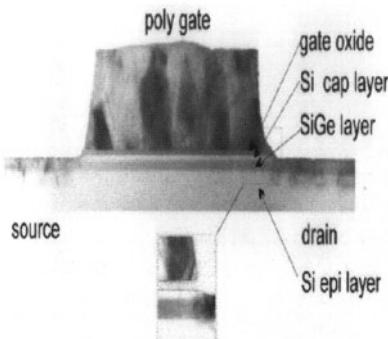


Fig. 18.3. TEM cross section of a SiGe p-channel MOSFET with a 10 nm SiGe layer (0.25 Ge) and 5 nm Si cap layer

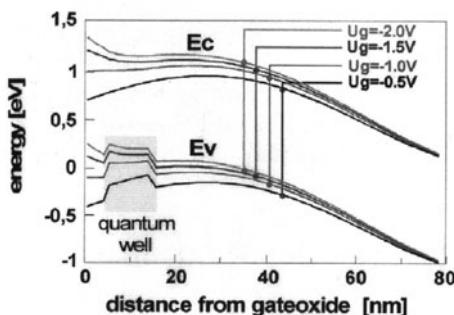


Fig. 18.4. Band structure of the SiGe p-channel transistor with a Ge content of 0.2 and n^+ poly-Si gate

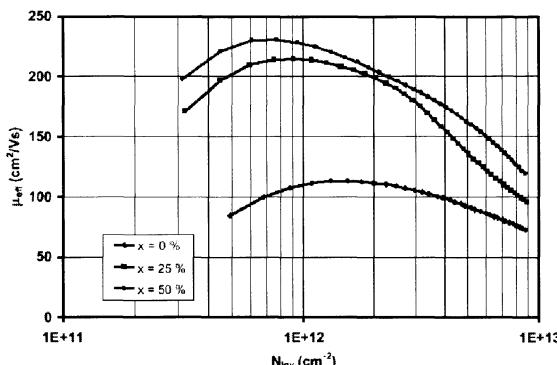


Fig. 18.5. Measured low-field mobilities in a p-channel transistor without Ge and with Ge contents of 0.25 and 0.5; $t_{\text{ox}} = 2.8 \text{ nm}$

When a gate voltage is applied, the holes are accumulated in the SiGe quantum well, which is under tensile strain owing to the smaller lattice constant of Si compared with SiGe. The mobility is enhanced because of the lower effective mass of the holes in SiGe and because the three degenerate valence bands are split, leading to less intervalley scattering. Moreover, the holes flow over an epitaxially grown interface, which is expected to be smoother than the Si/SiO₂ interface and therefore to result in less surface scattering.

The measured low-field mobilities with and without SiGe are depicted in Fig. 18.5. Compared with pure Si, with a peak hole mobility of about 110 cm²/Vs, the best values achieved are 210 cm²/Vs with a Ge content of 0.25 and 230 cm²/Vs with a Ge content of 0.5.

Higher Ge concentrations than 0.5 are difficult to achieve because of dislocations, which relax the stress in the SiGe layer during growth and subsequent high-temperature process steps.

Whereas the low-field mobility is a good indicator of the quality of the SiGe layer, transistors usually operate at much higher gate and drain voltages. Therefore, the saturation current is more interesting for practical applications. Owing to the high vertical and lateral electric fields, the mobility is greatly decreased. In Fig. 18.6, the saturation current characteristics of a SiGe p-channel transistor with a Ge content of 0.5 are shown for a gate length of 300 nm and compared with a pure Si device. An increase of the saturation current of 20–30% is achieved with SiGe at the same off current.

18.3.2 Strained Silicon

The best approach for the channel of the transistor will be to improve not only the hole mobility but also the electron mobility. This can be achieved with more complicated SiGe layer stacks and dual channels for electrons and holes [5]. A very CMOS-compatible approach is to use strained Si material with a surface channel. Here, the mobilities of both holes and electrons can

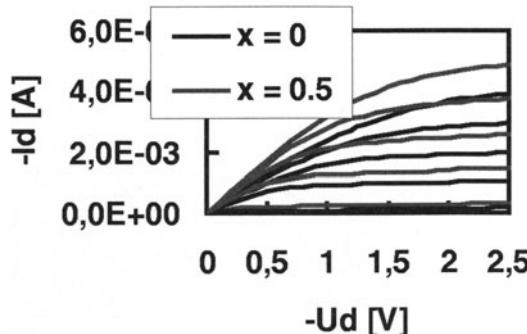


Fig. 18.6. Measured I - V characteristics of SiGe (Ge content 0.5) and Si p-channel transistor with $L = 300$ nm, $t_{ox} = 2.8$ nm, $V_g = 1.5, 1.3, 1.1, \dots$ V

be improved, depending on the strain in the Si layer. According to theoretical calculations [6], the hole mobility can be increased by up to a factor of 2.5 and the electron mobility by up to a factor of 2 for a high strain in the layer. The strain is created by a graded SiGe buffer layer, typically with a thickness of about 3 μ m and a Ge concentration of 20–30%. About half of the SiGe layer is graded with an increasing Ge content, and the rest is grown with a constant Ge concentration. Owing to the lattice mismatch, a large amount of dislocations are formed in the graded SiGe layer in order to relax the layer. In the upper part of the SiGe buffer layer, with a constant Ge concentration, the dislocation density should be as low as possible. Finally, on top, a thin Si layer with a thickness in the range of 10 to 20 nm is grown, which is now under tensile strain owing to the larger lattice of the SiGe substrate. Typically, a dislocation density in the Si in the range of 10^5 cm $^{-2}$ can be achieved today,

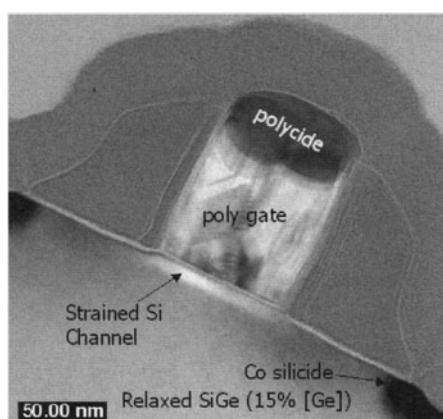


Fig. 18.7. TEM cross section of a 70 nm strained-Si MOSFET [7]

and process development with improved techniques to fabricate low-defect strained Si layers is going on.

A TEM cross section of a 70 nm strained-Si MOSFET [7] with a relaxed SiGe buffer layer with a 15% Ge content is shown in Fig. 18.7. The current flows at the $\text{SiO}_2/\text{strained-Si}$ interface. Measured low-field mobilities, in comparison with the universal Si mobility curve, yield an improvement of 70% for the n-channel transistor for Ge concentrations in the buffer layer of up to 20%, see Fig. 18.8. This corresponded to an increase in saturation current of 35% for the n-channel transistor [7].

For the p-channel transistor, little or no improvement was found for these Ge concentrations, but an increased mobility is expected for the holes at higher Ge concentrations [8] in the buffer layer.

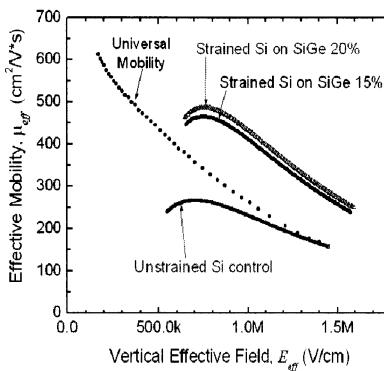


Fig. 18.8. Effective electron mobility for strained and unstrained Si [7]

18.3.3 Vertical Transistors

As shown in the ITRS Roadmap (see Fig. 18.2), the gate length of the transistors is much smaller than the corresponding technology node, which is defined as half the pitch of the critical gate layer. Therefore, lithography cannot provide these feature sizes and, special gate-trimming process steps are needed, such as an isotropic overetch of the gate or resist ashing of the gate mask.

A different approach is the vertical transistor, where short channel lengths can be processed without lithography. The three different main concepts for vertical transistors [9–11] are shown in Fig. 18.9.

The technologically simplest approach is the sidewall transistor. The current flows along a silicon sidewall defined by etching of a shallow trench, and is controlled by a poly-silicon spacer gate. The implantations for the source and drain and the wiring are very similar to those for planar devices. The bulk doping concentration of the transistor shown in Fig. 18.10 is $2 \times 10^{18} \text{ cm}^{-3}$.

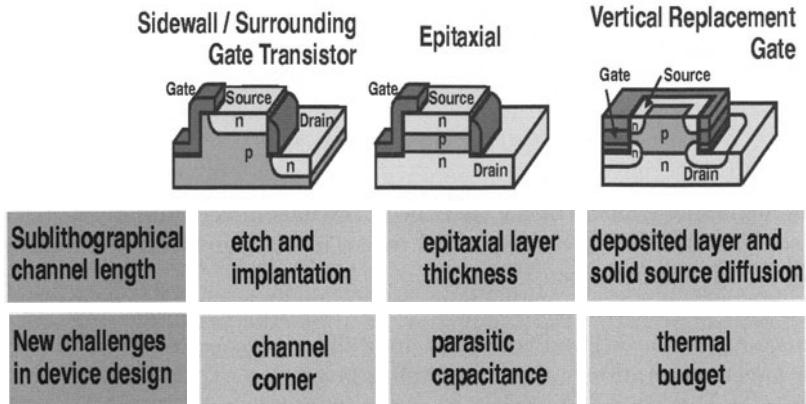


Fig. 18.9. Different architectures for vertical transistors, and the technology challenges

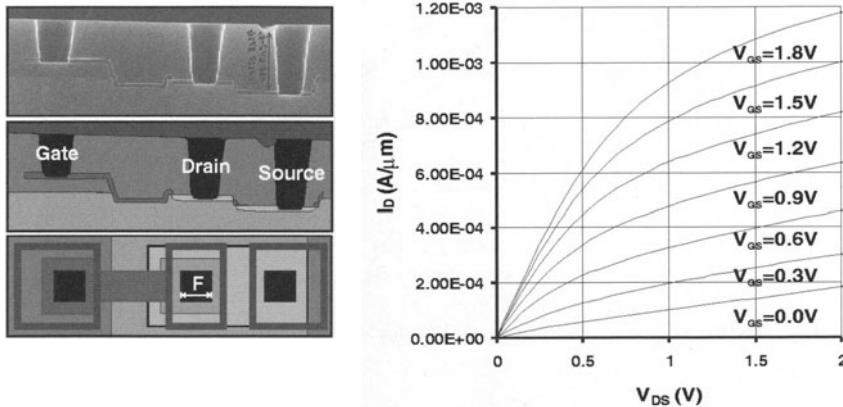


Fig. 18.10. Cross section and layout of a 50 nm vertical transistor, and measured I - V characteristics of an n-MOSFET ($t_{\text{ox}} = 3 \text{ nm}$, $N = 2 \times 10^{18} \text{ cm}^{-3}$)

At 1.2 V power supply voltage, a high oncurrent of $700 \mu\text{A}/\mu\text{m}$ was measured, but owing to short-channel effects the device cannot be turned off properly down to nA currents. Thus, higher doping concentrations will be needed for the channel.

Transistors with doping concentrations up to $1 \times 10^{19} \text{ cm}^{-3}$ have been investigated [9] (see Fig. 18.11). Now the device turns off properly at low drain voltages, but for higher drain voltages Zener tunneling currents from channel to drain appear and increase the off current again. Moreover, the on current is reduced, owing to the high channel doping concentration, as well as the subthreshold slope.

Therefore, new device concepts for MOSFETs without doping or with lower doping concentrations will be needed to overcome these obstacles.

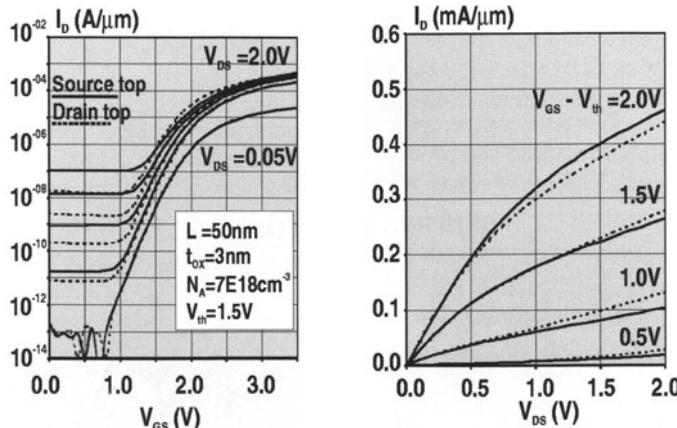


Fig. 18.11. Transfer and output characteristics of a 50 nm vertical n-MOSFET with a high channel doping concentration

18.3.4 Partially and Fully Depleted SOI

A promising approach to solving many of the device problems due to short-channel effects, shallow junctions, high doping concentrations and high junction capacitances is offered by silicon-on-insulator [12]. Detailed, material-related information is given in Chap. 8. Wafers of this material are now available in good quality at decreasing prices. Several semiconductor companies already offer SOI products, mainly for high-performance microprocessors or low-power applications. Here the thickness of the Si layer is in the range of 50 to 100 nm and the doping concentrations are similar to bulk devices. Therefore, this technology is classified as partially depleted SOI, because the width of the space charge layer is smaller than the Si layer thickness. Besides several advantages, there are some drawbacks related to the kink effect and to hysteresis effects caused by a charging of the floating Si body by avalanche multiplication. For short-channel transistors especially, thinner Si layers with fully depleted channels are of interest. If the Si thickness is reduced to 5 to 10 nm, the off current can reach very low values even with a low or undoped channel and can achieve a subthreshold slope close to the ideal value $kT \ln 10 = 60\text{ mV/dec}$, where k is the Boltzmann constant and T the temperature. Undoped channels are not feasible in bulk Si devices, because of punch-through from source to drain. Without channel doping, Zener tunneling currents are reduced, and also electrical parameter variations due to statistical fluctuations of the doping atoms. Moreover, the mobility of the charge carriers is higher in the channel.

In Fig. 18.12 a schematic cross section of a fully depleted SOI transistor [13] is shown. It has been derived from device simulations that a Si thickness of about one-quarter of the gate length is needed in order to achieve good device characteristics. Therefore, ultrathin Si layers are needed for transistors with a 25 nm gate length and below. To contact the source and drain regions

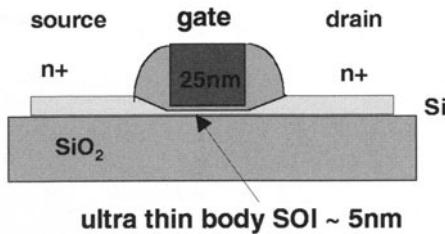


Fig. 18.12. Schematic cross section of a fully depleted SOI transistor and elevated source and drain regions

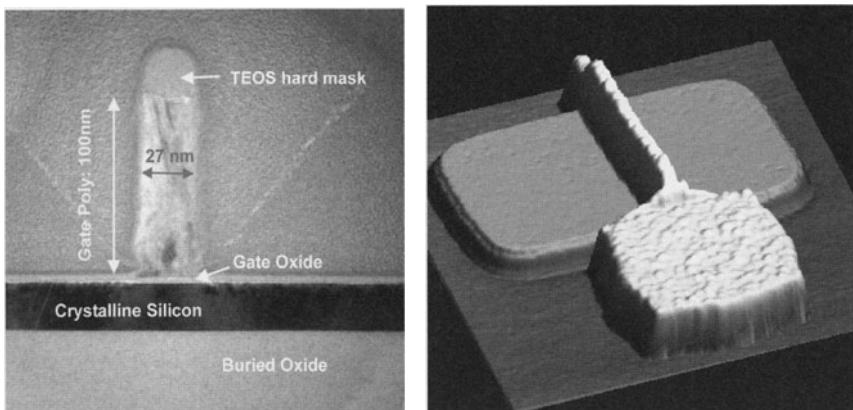


Fig. 18.13. Left, TEM cross section of an SOI MOSFET with 27 nm poly gate, 2.5 nm SiO_2 and 30 nm Si. Right, AFM micrograph of the SOI MOSFET after gate etch

and for silicidation, elevated source and drain regions are needed and have to be grown by Si epitaxy. Additionally, the threshold voltage has to be adjusted via the work function of the gate. Therefore, poly-Si has to be replaced by appropriate metal gates for n- and p-channels to achieve a threshold voltage of 200 to 300 mV.

In Fig. 18.13, an experimental realization of a thin-body SOI transistor with 27 nm gate length, 30 nm Si thickness, a buried-oxide thickness of 100 nm and a conventional poly-Si gate is shown. The gate was defined with e-beam lithography, and a mesa isolation with optimized spacers was used for the active area. Experimental I - V characteristics of n-channel SOI transistors with gate lengths down to 25 nm and with a 25 nm Si thickness are given in Fig. 18.14. In those devices, the off current was adjusted with a boron implantation of 10^{18} cm^{-3} and reaches $5 \text{ nA}/\mu\text{m}$ at $V_g = 0 \text{ V}$ for 50 nm gate length. Owing to the relatively thick Si body of 25 nm, short-channel effects increase for a 25 nm gate length, but the off current at $V_g = 0 \text{ V}$ is still below the ITRS 01 high-performance specification of $1 \mu\text{A}/\mu\text{m}$. For better

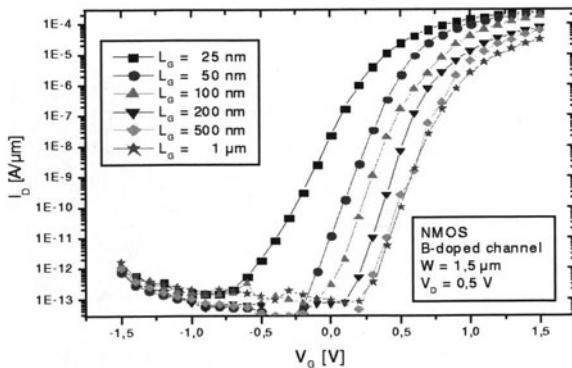


Fig. 18.14. Measured I - V characteristics of n-channel SOI MOSFETs down to 25 nm gate length with Si thickness of 25 nm, $t_{\text{ox}} = 2.5$ nm

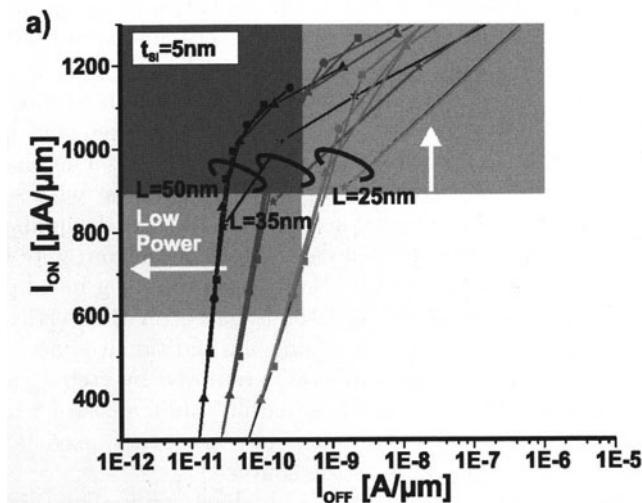


Fig. 18.15. Simulated I_{off} and I_{on} currents for ultrathin SOI transistors with 5 nm Si and $t_{\text{eq}} = 1$ nm

subthreshold slopes and lower off currents, much thinner Si layers are needed to turn off the channel in the bottom part also.

The expected performance for optimized SOI transistors with a 5 nm Si thickness and 1 nm equivalent oxide thickness, assuming a heavily nitrided gate oxide or a high- k dielectric, has been evaluated with device simulations using ATLAS. For gate lengths of 50, 35 and 25 nm the ITRS 01 specifications for high performance and low power can be reached, but the situation for an 18 nm gate length will become critical [14] (see Fig. 18.15).

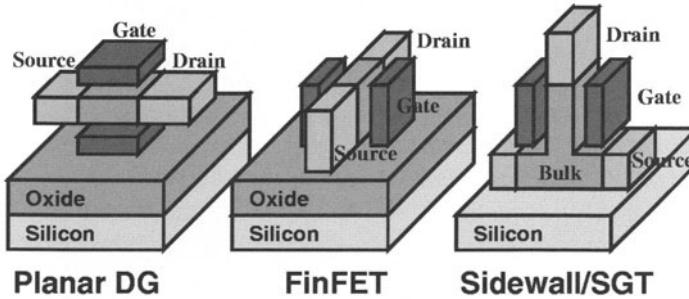


Fig. 18.16. Different architectures for double-gate transistors

18.3.5 Double-Gate Transistors

Further reduction of the channel length will need two or more gates for the control of the channel, together with thin Si layers, instead of only a single gate.

The advantage of the two gates is that they suppress the drain field much more effectively. Moreover, the Si thickness can be relaxed and the oncurrent is doubled, owing to the two transistors in parallel. The challenge for double-gate transistors will be to develop a reliable process with self-aligned gates, which is not easy to achieve, see Fig. 18.16. For the planar double-gate concept, the etching of a defined cavern for the bottom gate is difficult. Wafer-bonding techniques [15] and SON [16] are promising new approaches.

Another type of double-gate transistor is based on the vertical-sidewall transistor discussed in Chap. 18.3.3. Using an additional spacer as a self-aligned mask, most of the Si source region is removed by etching and only a thin Si ridge remains, which now forms a double-gate transistor with vertical current flow [9]. The current flows between the top contact, used as the drain, to the bottom contact, which acts as the source.

Recently the FinFET was proposed [17], which can realize this structure with relatively simple processing. The two gates are located vertically at the sidewalls of a thin Si channel, while the current flows laterally.

In Fig. 18.17, a FIB, focussed ion beam cross section of a small FinFET is shown. The fin and the gate layer were processed by e-beam lithography. The height of the fin is 50 nm and the width 15–25 nm; the buried-oxide thickness is 100 nm.

An SEM cross section through an enlarged FinFET processed by optical lithography [18] is shown in Fig. 18.18. The source and drain regions have been enhanced using selective Si epitaxy to lower the sheet resistance, and the top of the Si fin is shielded by a TEOS, tetra-ethyl-ortho-silicate/nitride cap layer.

The measured I – V characteristics of small n- and p-channel FinFETs are depicted in Fig. 18.19. The gate length for the n-channel FinFET is 20 nm, the fin width is 15 nm and the fin height is 50 nm. Despite a relaxed

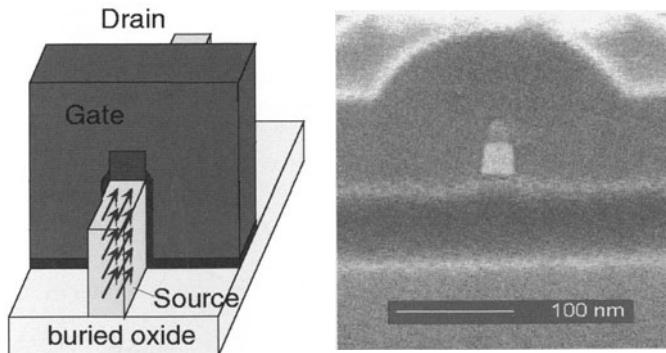


Fig. 18.17. FinFET-type [9] double-gate transistor with two channels in the SOI layer, and a device under processing after etching of a 50 nm high Si active region (transistor width)

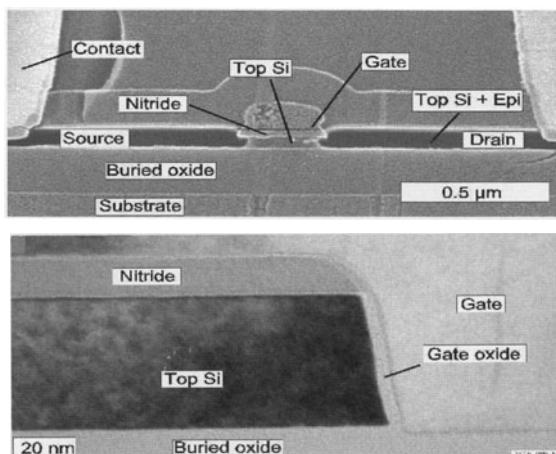


Fig. 18.18. TEM cross section of FinFET with elevated source and drain regions

gate oxide thickness of 3 nm, a high on current of $1.1 \text{ mA}/\mu\text{m}$ is achieved at $V_g - V_t = 1.1 \text{ V}$ with an undoped channel. Owing to the stronger diffusion of the boron-doped source and drain regions, the minimum gate length for the p-channel is 80 nm. This corresponds to about 40 nm effective channel length. Here the off current is in the range of $\text{pA}/\mu\text{m}$ and the on current reaches a remarkable $500 \text{ pA}/\mu\text{m}$.

A simulation of the expected subthreshold slope, as a measure of short-channel effects for single- and double-gate MOSFETs on SOI, is depicted in Fig. 18.20. For channel lengths above 50 nm both transistors exhibit the ideal value of 60 mV/dec. The single-gate SOI transistor starts to degrade at 25 nm gate length, whereas the double-gate transistor will operate with good performance down to 10 nm, assuming a Si thickness of 5 nm for both devices.

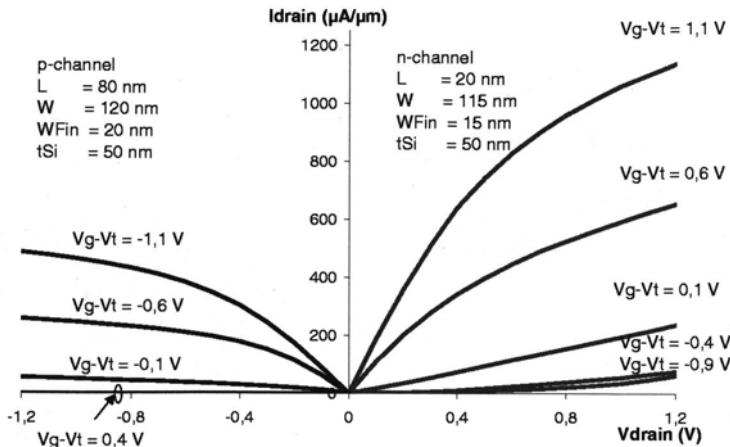


Fig. 18.19. Measured I - V characteristics for n- and p-channel FinFETs with 20 nm and 80 nm gate length, $t_{ox} = 3 \text{ nm}$

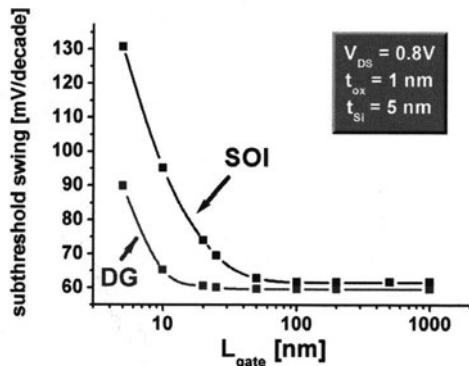


Fig. 18.20. ATLAS simulation of double-gate and single-gate SOI n-MOS transistors with $t_{ox} = 1 \text{ nm}$ and Si thickness 5 nm

18.4 FinFET Memory Cell

Today, the storage capacity of integrated memory cells such as DRAM (dynamic random access memory) [19] and nonvolatile Flash EEPROM (electrically erasable and programmable read-only memory) [20] is in the 256 Mbit–1 Gbit range. And the demand for even higher densities is still increasing, in particular for many nonvolatile data and code storage applications. Currently, the most widely used memory cell in a flash EEPROM consists of a transistor with a floating gate [20] or a storage dielectric sandwiched between the gate electrode and the channel region [21]. A small amount of charge is

transferred to this storage electrode and kept persistently, and is easily read out by a shift in the I - V characteristics of the transistor.

As for transistors, the scaling properties of such devices have been excellent in the past but will be more challenging in the coming years. Nevertheless, this cell type remains very promising even for smaller transistor sizes, because the threshold shift increases with decreasing gate capacitance. At very small dimensions of ~ 20 nm the storage region starts to act effectively as a quantum dot, and a few electrons are sufficient to induce a threshold shift of 1 V according to $\Delta V_t = ne/C$, where n is the number of electrons. In principle, even a single electron can be sensed at room temperature. The real challenge for these small devices will be retention time, since storing information with only a few electrons necessitates very reliable isolation dielectrics. Furthermore, new device architectures such as the double gate may be required that allow better electrostatic channel control and screening of the stored charges from parasitic voltages.

In Fig. 18.21, left side, a schematic cross section of a FinFET double gate transistor is shown, with a multilayer gate dielectric at the sidewalls as the storage element. The planar top region of the fin is shielded by an oxide layer. On the right-hand side, a TEM cross section of a memory structure realized using e-beam lithography, with a Si fin of 32 nm width and 38 nm height, is shown. The multilayer dielectric consists of 5 nm SiO_2 , 5 nm Si_3N_4 and 5 nm SiO_2 . The gate electrode is poly-Si with a gate length of 50 nm. Owing to the double gates and the thin fin, short-channel effects are reduced compared with planar single-gate transistors.

The electrical principle has been verified experimentally using an enlarged FinFET processed by optical lithography and with a nitride trapping layer. When a gate voltage of +4.5 V (write) is applied, electrons are injected into the nitride trapping layer. Therefore the I - V curves are shifted to positive voltages. Applying a gate voltage of -45 V (erase) pushes the electrons out of

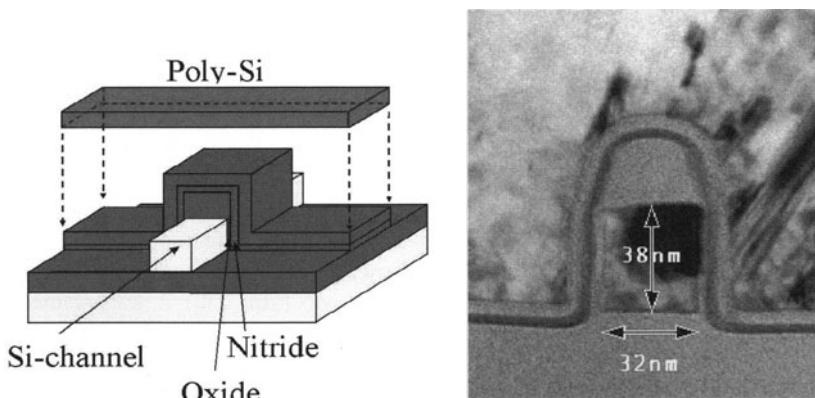


Fig. 18.21. Schematic FinFET memory transistor and experimentally realized structure with Si fin, multilayer dielectric and poly-Si gate

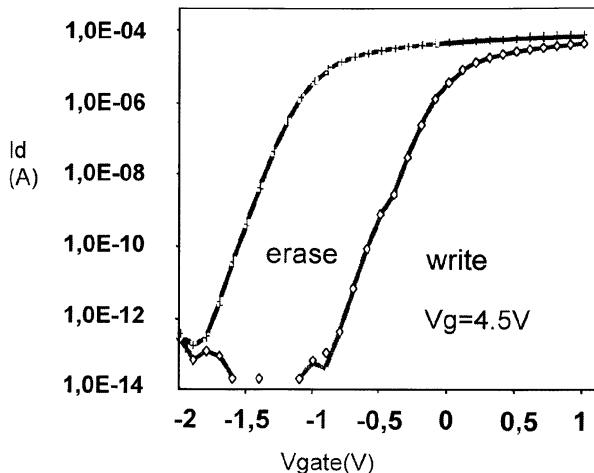


Fig. 18.22. Measured I - V shifts in a FinFET quantum dot memory device with a nitride trapping layer

the nitride layer and shifts the I - V curve back to negative voltages. A shift of about 1 V was obtained at a gate length of 300 nm and a 50 nm fin height (see Fig. 18.22).

18.5 Limits of Si MOSFETs

In 1972, the first microprocessor was produced in high volume, with 5 μm feature sizes and 4000 transistors. Soon, rapid progress in integration density was achieved [22]. Very early, an extensive discussion started about the limits and the end of scaling of microelectronics. In 1984, the limit was considered to be 0.2–0.4 μm [23]. Today, CMOS with gate lengths down to 70 nm is in production, and in research laboratories functional 14 nm bulk Si transistors [24] and even ultrathin SOI transistors [25] with a 6 nm gate length have already been demonstrated. However, their electrical performance still has to be improved before production to obtain a benefit from scaling. This improvement is expected from high- k dielectrics, metal gates, strained Si and multiple-gate structures, as discussed previously.

A simplified drift-diffusion simulation using the device simulator ATLAS (see Fig. 18.23), shows the achievable performance of a 6 nm double-gate transistor assuming a gate oxide thickness of 1 nm and a very aggressive Si thickness of only 2 nm. At 0.6 V power supply voltage, an on current of 1.2 mA/ μm seems to be possible, and the device can be turned off well.

But at these dimensions, quantum mechanical effects should be taken into account. The most important are direct tunneling currents through the thin gate dielectric, tunneling currents from the valence band of the channel directly to the drain and, for very short channels, tunneling currents consisting

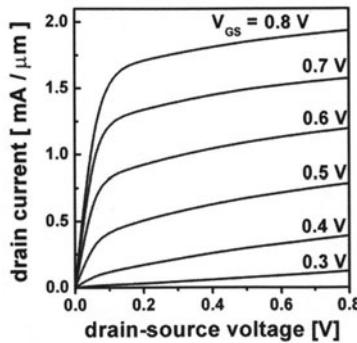


Fig. 18.23. Simplified ATLAS device simulation of a 6 nm double-gate transistor using the drift-diffusion model, $t_{\text{ox}} = 1 \text{ nm}$, $t_{\text{si}} = 2 \text{ nm}$

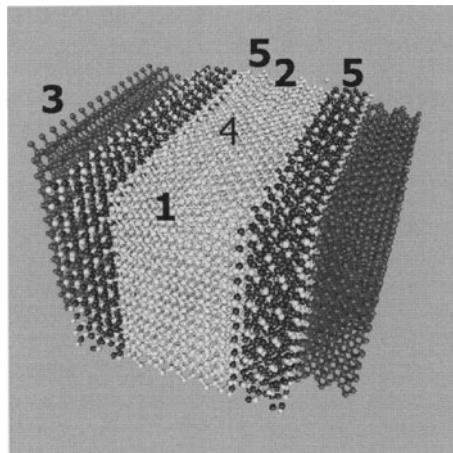


Fig. 18.24. Atomistic view of a nanoscale FinFET double-gate transistor: (1) source, (2) drain, (3) gate, (4) silicon, (5) SiO_2

of conduction band electrons that tunnel from the source to the drain through the potential barrier of the channel. Also, owing to the wave function of the electron, the density does not have its maximum at the interface between the Si and the dielectric, and the quantization of the energy states in the inversion layer gives corrections to the transistor current. Finally, ballistic carrier transport should be taken into account in the sub-50 nm regime.

Assuming source-drain tunneling to be the ultimate limitation of MOS-FETs with decreasing gate length, an atomistic simulation approach [26] based on the tight-binding method was used for simulation. A double-gate FinFET with a thickness of 2.5 nm, corresponding to five Si lattice constants, is depicted in Fig. 18.24 in an atomistic view. The 1 nm thick gate dielectric consists of 6 Si-O bond lengths at the two sidewalls and is covered by the gate electrode. The current flows from the source to the drain either thermally

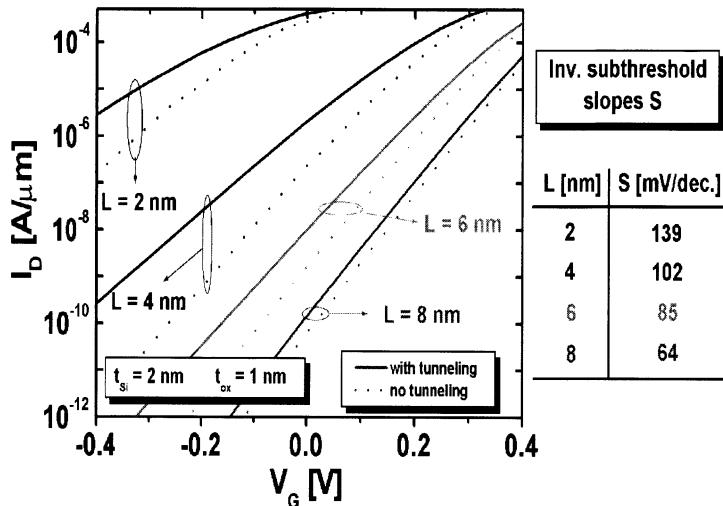


Fig. 18.25. Simulated drift-diffusion and tunnel currents using the tight-binding method for a double-gate MOSFET with 2 nm Si thickness and 1 nm SiO_2 dielectric

across the potential barrier near the source or by direct tunneling through the potential barrier. The simulated current flowing from the source to the drain with and without source-drain tunneling is given in Fig. 18.25 for various gate lengths; the subthreshold slopes have been extracted and are given on the right-hand side of the figure.

The tunnel currents do not greatly deteriorate the transistor characteristics down to 6 nm, and even for gate lengths of 2–4 nm the current can be modulated with the gate. But the limit of MOSFET miniaturization will be in this range, corresponding to 4 Si lattice constants. Remarkably, this is in the range of the diameter of a carbon nanotube, of about 1.2 nm, or the length of a small molecule.

18.6 Emerging Devices

18.6.1 Single-Electron Transistors

While in nanoscale MOSFETs pn junctions are used for the isolation of the source and drain regions and many charge carriers flow in the channel, a similar device structure leads to single-electron devices [27]. These are also three-terminal devices with a source, drain and gate, where the gate controls the current electrostatically as in MOSFETs. The substrate can be silicon or a metal on an insulator, but the pn junctions are replaced by high-resistivity tunnel junctions (see Fig. 18.26). The channel region has to be made extremely small; for room temperature operation a length and a width of 1–2 nm are needed. This yields capacitances in the range of 0.1 aF, corresponding to a Coulomb energy of ~ 1 eV for one electron to enter the channel.

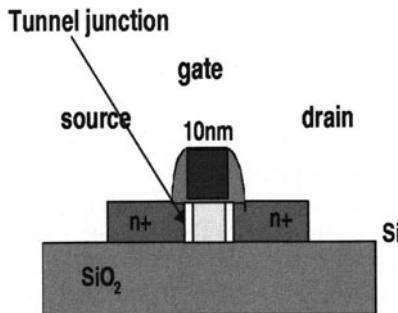


Fig. 18.26. Single-electron transistor on SOI with tunnel junctions

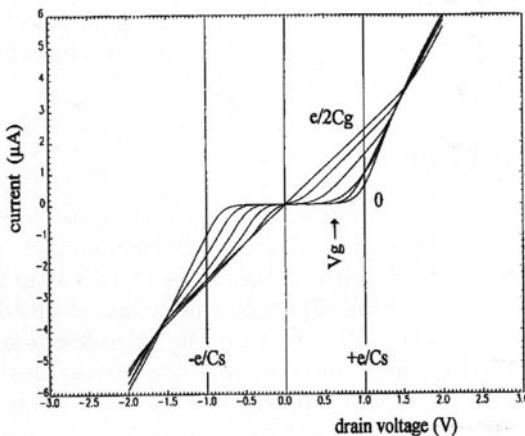


Fig. 18.27. Simulated output characteristics of single-electron transistor at 385 K ($R_t = 100 \text{ k}\Omega$, $C_g = 0.032 \text{ aF}$, $C_t = 0.064 \text{ aF}$)

After one additional electron is in the channel, the next electron cannot enter from the source for a given gate voltage, owing to the negative charge, until the first electron has tunneled to the drain.

So, at any time, one additional electron is always in the channel. As can be seen in the simulation of the I - V characteristics [28] in Fig. 18.27, the current is low for drain voltages below $e/C - s = 1 \text{ V}$ and can be turned on with the gate voltage in accordance with the orthodox theory. A maximum on current is achieved at a gate voltage of $e/2C_g$. C_g corresponds to the gate/channel capacitance, and C_s is the sum of the capacitances of the two junctions C_t and C_g .

The maximum current depends on the tunnel resistance, which must be larger than the Klitzing resistance $h/e^2 = 26 \text{ k}\Omega$. In the simulation $100 \text{ k}\Omega$ was assumed, so the current is in the range of several μA .

In contrast to MOSFETs, the transfer characteristic is periodic with gate voltage, but it behaves similarly for half a period of the gate voltage. The

subthreshold slope obtained in a simulation of an optimized single electron transistor is about 80 mV/dec, as in a MOSFET. This is related to the energy distribution of the charge carriers and can be influenced only by temperature.

Owing to the fact that for SET, single electron transistor it is much more difficult to optimize the ratio of the gate/channel to the drain/channel capacitance, an even stronger dependence on drain voltage than in a MOSFET is observed. Note also that an extremely small capacitance of 0.032 aF was assumed in the simulation. Because only one electron flows in the channel, these devices are also very sensitive to offset charges and are much slower in switching speed than CMOS devices are. Moreover, the extreme requirements on the technology if sub-1 nm reproducibility is to be obtained are far away from today's processing capabilities. Nevertheless, it is of fundamental interest to operate a device with a charge of one electron in the channel, and Coulomb blockade will also become an issue for nanoscale MOSFETs that use not one but only a few electrons in the channel, and when storage of a few electrons is used in a memory device.

18.6.2 Molecular Devices

Another approach to reaching nanometer feature sizes is to use molecules as electronic devices. The advantage of the molecular approach will be that the electrically active structures, with feature sizes of 1 to 10 nm, will not need extreme lithography and etching. The molecules will be formed and deposited by chemical means and will probably exhibit no tolerances due to processing. Resistors built from rotaxane- and catenane-type molecules [29] that show a different conductivity after an applied voltage pulse have been reported, in a 64 bit resistive cross-bar memory cell array. Also, transistors where the current flowing through the molecules can be modulated with a gate seem possible. A first step to a hybrid molecular-Si-CMOS memory array [30] is shown schematically in Fig. 18.28. The x and y wires of the cell array will be fabricated using conventional CMOS interconnect technology with some process modifications such as Au electrodes. The address, sense and input/output circuits will be in CMOS technology. But the storage is based on molecules, which are located in a monomolecular film at the intersections of the word and bit lines. The resistance of the molecules, and thus the information, is stored in a nonvolatile way and can be changed by appropriate voltages on the word and bit lines. The current related to the resistance will be sensed by a CMOS sense amplifier. This could lead to cheap, very dense nonvolatile memory elements and will be very helpful for further scaling of CMOS memories.

18.6.3 Carbon Nanotubes

The biggest progress in emerging devices has been achieved with carbon nanotubes in the last few years. Carbon nanotubes consist of a perfect crystal lattice with a folded hexagonal structure of C atoms. Depending on the

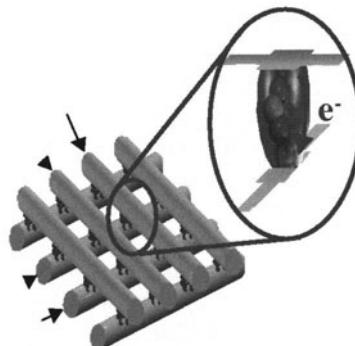


Fig. 18.28. Schematic approach to a molecular memory array in combination with CMOS

orientation, the nanotubes are metallic conductors or semiconductors. The diameter of a single-wall carbon nanotube is 1.4 nm and the length is typically several μm . Also, the growth of thicker, multiwall nanotubes is possible. Similarly to Si MOSFETs, carbon nanotube transistors have been built, with metallic regions for the source and drain, a gate dielectric, and a gate electrode. However, the channel region consists of a carbon nanotube instead of silicon. In [31] a ratio of the on and off currents of about five orders of magnitude and a subthreshold slope of 170 mV/dec have been achieved, similar to the values achieved in Si MOSFETs. Owing to the small width of the nanotube, the on current is in the range of μA . But if the current is normalized to a width of μm as for Si MOSFETs, the on current reaches several $\text{mA}/\mu\text{m}$ and could even be two to three times than in a Si MOSFET owing to the perfect carbon lattice. Now the challenge will be to develop a reproducible technology for the growth of single or multiple carbon nanotubes in a self-assembling way on the source and drain regions. Recent progress, basic technology and the physics of carbon nanotubes will be discussed elsewhere in this book in more detail (Chap. 23).

18.7 Perspectives

Focusing on the Si MOSFET, it seems very likely that the scaling of Si CMOS will continue down to the 22 nm node in the year 2016, in accordance with the ITRS Roadmap. Considering only device performance, CMOS FETs can operate well below 10 nm gate length. Eventually, source-drain tunneling at channel lengths of 2 to 4 nm will increase the off currents of the transistors to unacceptably high values and will stop further miniaturization. But this is really far in the future, and therefore it is predicted that the technological progress in microelectronics due to scaling will continue in the era of nanoelectronics.

In this positive scenario, it is assumed that lithography tools such as extreme ultraviolet, parallel e-beam or ion beam lithography will be available for such small feature sizes. Also, the critical wiring issue of the devices has not been discussed as the real limiting problem here. This is because designers use multilevel metallization with different sizes and lengths of wires. As long as short distances not exceeding several μm can be used for the fine-line wires, the transistor always gives the dominant contribution to the delay time. In addition, another possible roadblock, the increasing power dissipation, has not been discussed. Here, circuit and system solutions are needed; the improvements available from technology seem to be limited.

Coming back to the devices, bulk CMOS may run into performance constraints at the 45 nm CMOS generation and below. But thin-film SOI can probably take over and will allow further downscaling. Finally, double-gate or multigate transistors with low-doped channels are considered to be the best nanodevices with respect to I_{off} , I_{on} , switching speed and integration density. They can approach the regime of single-electron transistors, which need a size of 1 nm for room temperature operation, and of single-wall carbon nanotubes with a diameter of 1.2 nm. Today, Si CMOS is the only realistic approach for very large-scale integrated logic and memory circuits. No other devices with better performance and manufacturability are in sight. Thus, the end of the roadmap seems not to be in sight at least for the next 25 years.

Acknowledgments

The author would like to thank his colleagues in the Department of Nano-devices M. Alba, D. Alvarez, L. Dreeskornfeld, J. Hartwich, F. Hofmann, G. Ilicali, J. Kretz, E. Landgraf, R.J. Luyken, E. Rutkowski, W. Rösner, T. Schulz, M. Specht and M. Staedele for their contributions, and, especially, K. Rim for Figs. 18.7 and 18.8.

References

1. *International Technology Roadmap for Semiconductors 2001*, SEMATEC, Austin, TX (2001)
2. J.T. Wallmark: Fundamental physical limitations in integrated electronic circuits. Solid State Devices, 1974 Conference Series **25**, 133 (1975)
3. R.W. Keyes: Physical limits in digital electronics. Proceedings of the IEEE 63 (1975) No. 5, p. 740
4. L. Risch, H. Fischer, F. Hofmann, H. Schäfer, M. Eller, T. Aeugle: Fabrication and electrical characterisation of Si/SiGe p-channel MOSFETs with a delta doped boron layer. Proceedings of the 26th European Solid State Device Research Conference (ESSDERC) (1996) p. 465
5. T. Mizuno, N. Sugiyama, T. Tezuka, T. Numata, T. Maeda, S. Takagi: Design for scaled thin film strained-SOI CMOS devices with higher carrier mobility. International Electron Devices Meeting, Technical Digest (2002) p. 31

6. R. Oberhuber, G. Zandler, P. Vogl: Subband structure and mobility of two-dimensional holes in strained Si/SiGe MOSFET's. *Physical Review B* **58** (15), 9941 (1998)
7. K. Rim, S. Koester, M. Hargrove, J. Chu, P.M. Mooney, J. Ott, T. Kanarsky, P. Ronsheim, M. Ieong, A. Grill, H.-S.P. Wong: Strained Si NMOSFETs for high performance CMOS technology. *Symposium on VLSI Technology* (2001) p. 59
8. J.L. Hoyt, H.M. Nayfeh, S. Eguchi, I. Aberg, G. Xia, T. Drake, E.A. Fitzgerald, D.A. Antoniadis: Strained silicon MOSFET technology. *International Electron Devices Meeting, Technical Digest* (2002) p. 23
9. T. Schulz, W. Rösner, L. Risch, U. Langmann: 50 nm vertical sidewall transistors with high doping concentrations. *International Electron Devices Meeting, Technical Digest* (2000) p. 61
10. L. Risch, W.H. Krautschneider, F. Hofmann, H. Schäfer: Vertical MOS transistors with 70nm channel length. *Proceedings of the 25th European Solid State Device Research Conference (ESSDERC)* (1995) p. 101
11. J.M. Hergenrother, D. Monroe, F.P. Klemens, A. Kornblit, G.R. Weber, W.M. Mansfield, M.R. Baker, F.H. Baumann, K.J. Bolan, J.E. Bower, N.A. Ciampa, R.A. Cirelli, J.I. Colonell, D.J. Eaglesham, J. Frackoviak, H.J. Gossmann, M.L. Green, S.J. Hillenius, C.A. King, R.N. Kleinman, W.Y.-C. Lai, J.T.-C. Lee, R.C. Liu, H.L. Maynard, M.D. Morris, S.-H. Oh, C.-S. Pai, C.S. Rafferty, J.M. Rosamilia, T.W. Sorch, H.-H. Vuong: The vertical replacement-gate (VRG) MOSFET: a 50 nm vertical MOSFET with lithography-independent gate length. *International Electron Devices Meeting, Technical Digest* (2000) p. 75
12. J.P. Colinge: *SOI Technology: Materials to VLSI*, 2nd edn (Kluwer, Boston, MA 1997)
13. L. Dreeskornfeld, J. Hartwich, E. Landgraf, R.J. Luyken, W. Rösner, T. Schulz, M. Städele, D. Schmitt-Landsiedel, L. Risch: Comparison of partially and fully depleted SOI transistors down to the sub-50nm gate length regime. *The Electrochem. Soc. Proc.: SOI Technology and Devices XI*, vol. 2003-05 (2003), p. 361
14. R.J. Luyken, M. Städele, W. Rösner, T. Schulz, J. Hartwich, L. Dreeskornfeld, L. Risch: Perspectives of fully-depleted SOI transistors down to 20 nm gate length. *Proceedings of 2002 IEEE International SOI Conference 10/02*, p. 137
15. H.S.P. Wong: Beyond the conventional MOSFET. *Proceedings of ESSDERC* (2001) p. 69
16. S. Monfray, T. Skotnicki, B. Tavel, Y. Morand, S. Descombes, A. Talbot, D. Dutartre, C. Jenny, P. Mazoyer, R. Palla, F. Leverd, Y. Le Fric, R. Pantel, M. Haond, C. Charbuillet, C. Vizioz, D. Louis, N. Buffet: SON (silicon-on-nothing) P-MOSFETs with totally silicided (CoSi_2) polysilicon on 5 nm-thick Si-films: the simplest way to integration of metal gates on thin FD channels. *International Electron Devices Meeting, Technical Digest* (2002) p. 263
17. X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. King, J. Bokor, C. Hu: Sub 50 nm FinFET: PMOS. *International Electron Devices Meeting, Technical Digest* (1999) p. 67
18. E. Landgraf, W. Rösner, R.J. Luyken: High on current in quasi double gate transistors with undoped channel region. *Proceedings of the 31th European Solid State Device Research Conference (ESSDERC)* (2001) p. 271

19. H. Seidl, M. Gutsche, U. Schroeder, A. Birner, T. Hecht, S. Jakschik, J. Luetzen, M. Kerber, S. Kudelka, T. Popp, A. Orth, H. Reisinger, A. Saenger, K. Schupke, B. Sell: "A fully integrated Al_2O_3 trench capacitor DRAM for sub-100 nm technology. International Electron Devices Meeting, Technical Digest (2002) p. 839
20. J.D. Choi, S.-S. Cho, Y.-S. Yim, J.-D. Lee, H.-S. Kim, K.-J. Joo, S.-H. Hur, H.-S. Im, J. Kim, J.-W. Lee, K. Ill Seo, M.-S. Kang, K.-H. Kim, J.-L. Nam, K.-C. Park, M.-Y. Lee: Highly manufacturable 1 Gb Nand Flash using 0.12- μm process technology. International Electron Devices Meeting, Technical Digest (2001) p. 211
21. B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, D. Finzi: NROM: A novel localized trapping, 2-bit nonvolatile memory cell. *IEEE Electron Device Letters* **21**, 543 (2000)
22. G.E. Moore: Progres in Digital Integrated Electronics. International Electron Devices Meeting, Technical Digest (1975) p. 11
23. J.D. Meindl: "Ultra large scale integration. *IEEE Transactions in Electron Devices* **ED-31**, No. 11, 1555 (1984)
24. A. Hokazono, K. Ohuchi, M. Takayanagi, Y. Watanabe, S. Magoshi, Y. Kato, T. Shimizu, S. Mori, H. Oguma, T. Sasaki, H. Yoshimura, K. Miyano, N. Yasutake, H. Suto, K. Adachi, H. Fukui, T. Watanabe, N. Tamaoki, Y. Toyoshima, H. Ishiuchi: 14 nm gate length CMOSFETs utilizing low thermal budget process with poly-SiGe and Ni salicide. International Electron Devices Meeting, Technical Digest (2002) p. 639
25. B. Doris, M. Jeong, T. Kanarsky, Y. Zhang, R.A. Roy, O. Dokumaci, Z. Ren, F.-F. Jamin, L. Shi, W. Natzle, H.-J. Huang, J. Mezzapelle, A. Mocuta, S. Womack, M. Gribelyuk, E.C. Jones, R.J. Miller, H.-S.P. Wong, W. Haensch: Extreme scaling with ultra-thin Si channel MOSFETs. International Electron Devices Meeting, Technical Digest (2002) p. 267
26. M. Staedele: Influence of source-drain tunneling on the subthreshold behavior of sub-10 nm double-gate MOSFETs. *Proceedings of ESSDERC* (2002) p. 135
27. Y. Takahashi, H. Namatsu, K. Kurihara, K. Iwdate, M. Nagase, K. Murase: Size dependence of the characteristics of Si single electron transistors on SIMOX substrates. *IEEE Transactions on Electron Devices* **43**, No. 8, 1213 (1996)
28. W. Rösner, F. Hofmann, T. Vogelsang, L. Risch: Simulation of single electron circuits. *Microelectronic Engineering* **27** (1-4), 55 (1995)
29. C.P. Collier, E.W. Wong, M. Belohradský, F.M. Raymo, J.F. Stoddart, P.J. Kuekes, R.S. Williams, J.R. Heath: Electronically configurable molecular-based logic gates. *Science* **285**, 391 (1999)
30. R.J. Luyken, F. Hofmann: Concepts for hybrid CMOS-molecular non-volatile memories. *Nanotechnology* **14** (2), 273 (2003)
31. P. Avouris, J. Appenzeller, V. Derycke, R. Martel, S. Wind: Carbon nanotube electronics. International Electron Devices Meeting, Technical Digest (2002) p. 281

19 Lithography for Silicon Nanotechnology

J. Kretz

19.1 Introduction

Lithography is one of the key methods of silicon device technology since it defines the structure of the silicon devices. However, at the same time, it is one of the most demanding topics because scaling is the mainspring of the advance of silicon devices, underpinning the rapid increase in performance during the last three decades (Fig. 19.1). In lithography, the structures are “printed” into an auxiliary layer, called the resist, which serves as a masking layer for a further structuring of the underlying layers and is removed after use. Optical lithography, where the structures to be printed are written onto a glass blank and transferred to the photosensitive resist by projection, is commonly used in industrial fabrication. Although the end of optical lithography has been predicted for more than ten years, this technique has been improved continuously and is currently serving its purpose very well. Recent enhancements of optical lithographic techniques are described below.

The expected end of optical lithography has led to ongoing research into new lithographic techniques for the post-optical-lithography era, called *next-generation lithographies* (NGLs). From the number of possible techniques under investigation a few years ago, the options have been narrowed down and at present only *extreme ultraviolet* (EUV) lithography is considered to be a worthy successor to optical lithography that guarantees the necessary throughput. Electron beam methods, which are serial, direct writing processes, are already established for the production of masks for optical and EUV lithography. Extensive knowledge of this technology has resulted in its use for the production of devices for low-volume production ASIC, application specific integrated circuits, prototyping and research. Parallel-writing strategies with many electron beams are still under discussion and are subject to examination to determine whether they are suitable as NGLs. Finally, there are structuring methods such as nanoimprint and proximal-probe lithography, where research groups have shown exciting results, pertaining mainly to ultrahigh resolution. Even though the majority of these techniques have elementary difficulties, in particular with overlay and throughput, it is worthwhile to keep track of the ongoing progress of these methods, since even bottom-up approaches (in contrast to the conventional top-down techniques of standard lithography) promise to deliver the smallest structures, even down to the atomic scale. All of these lithographic methods are described briefly

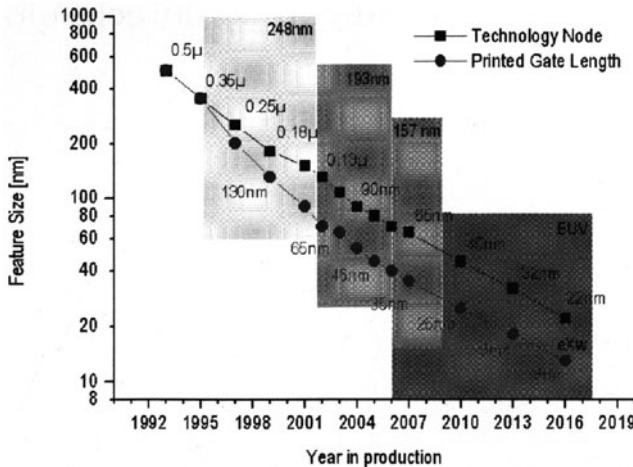


Fig. 19.1. The ITRS lithography roadmap [1], showing minimum feature sizes and the corresponding lithographic technique. The *square points* indicate the technology node (half of the pitch for printing dense structures), whereas the *circular data points* show the printed gate length in the resist (single lines) after the lithographic step. The physical gate length of the transistor, i.e. the gate length after etch, has normally been 40% smaller than the printed gate length from the year 2000 on

below and analyzed according to the role they could play in nanostructure fabrication.

19.2 Optical Lithography

Optical lithography is currently the dominant exposure process. In addition to the existence of mask aligners for research and development, where mask blanks are brought into proximity or direct contact with a resist-coated substrate and the structures of the mask are exposed directly onto the wafer, steppers are used in high-volume production. In a stepper, projection optics reduce the aerial image of a single chip onto the substrate, which is scanned stepwise to define multiple chips on the wafer. The reduction factor of the projection lens is usually four to five, which relaxes the specifications for the pattern mask, called the reticle. The high-precision stepping of the stepper is accomplished with a laser interferometer table, which enables a placement accuracy of better than 50 nm. This is necessary since the demands on overlay accuracy are currently smaller than 35 nm and will decrease to less than 10 nm by 2016.

The key equations for the optical system of a stepper are given by

$$R = k_1 \cdot \frac{\lambda}{\text{NA}}, \quad (19.1)$$

$$\text{DOF} = k_2 \cdot \frac{\lambda}{\text{NA}^2}, \quad (19.2)$$

where R is the feature size, **DOF** is the depth of focus, λ is the wavelength of the light used, **NA** is the numerical aperture of the lens system (aperture angle), and k_1 and k_2 are process parameters. Equation (19.1) follows from the Rayleigh criterion. For an aerial image $k_1 = 0.5$, but the exact value is determined by the process conditions such as resist behavior and illumination mode. Equation (19.2) gives the depth of focus (**DOF**), which can also be derived from the Rayleigh criterion, whereby the prefactor k_2 is process-dependent, similarly to k_1 .

To achieve smaller line widths, the direct approach is to reduce the wavelength λ used. In Fig. 19.1 it can be seen that wavelengths of 248 nm (KrF excimer laser), 197 nm (ArF) and 157 nm (F_2) are used for optical lithography. A change of the wavelength means a new machine generation, whose cost increases by a factor of 10 per generation, mainly as a result of the new optical system for the specific wavelength. Since an improvement of the wavelength by 20% is insufficient on its own, an increase in **NA** and a decrease in k_1 are also aimed for. When **NA** is increased a trade-off has to be found because **DOF** depends quadratically on **NA**. At present **NA** typically ranges from 0.6 to 0.75. Finally, k_1 can also be decreased by process improvements, e.g. resist optimization, reticle engineering or off-axis illumination.

Off-axis illumination is achieved by the introduction of an annular or a quadrupole aperture into the illumination system, permitting higher diffraction orders to enter the objective, resulting in an increase of the resolution. Theoretically k_1 can be reduced to a minimum value of 0.25 with off-axis illumination, but it is normally considerably larger.

The reticle itself is a quartz blank with an electron-beam-patterned chromium absorber layer. Usually the chromium absorbs the light, yielding an intensity gradient at the edge of the structure. Reticle engineering is applied to create steeper intensity edges in the structures in the aerial image by modifying the chromium-on-glass mask. One may differentiate between attenuated (ATN), alternating (ALT) and chromiumless (CRL) phase masks. In all cases, an additional phase shift of the wave at the boundaries of structures is accomplished, either by replacing the chromium by a 180° phase-shifting material (ATN) or by thinning the glass substrate in places where the chromium has been removed in order to obtain a phase shift (ALT). The option of creating the phase shift solely by etching into the glass blank to generate a locally steep intensity profile leads to a CRL mask.

Finally, the optical proximity effect manifests itself by a difference in the critical dimensions (CD; the present target value is $3\sigma < 5$ nm) depending on surrounding structures, end-of-line shortening depending on the feature size, and a corner-rounding effect. All these effects can be responded to with different optical-proximity-effect corrections (OPC), e.g. feature biasing (thinning of isolated lines), the addition of subresolution features (scatter bar mode), line length biasing (compensation for line shortening, i.e. line mode), and the addition of hammerheads, serifs or segments to corners of structures (see

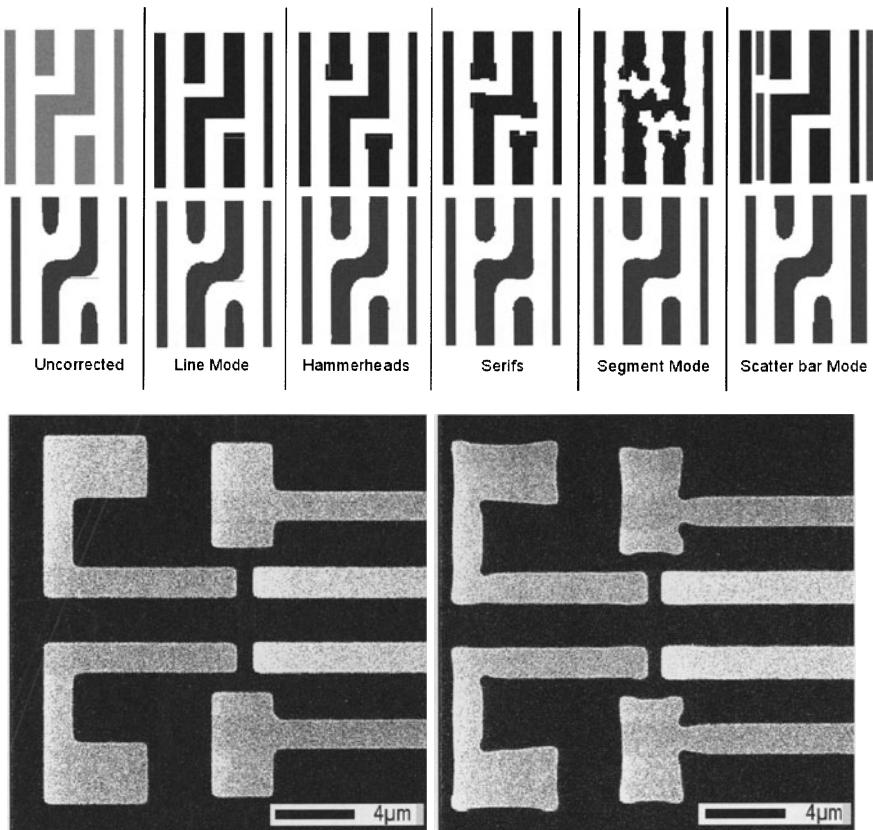


Fig. 19.2. Optical-proximity-correction methods used in state-of-the-art optical reticles. In the *upper picture* the structure on the reticle can be seen (*pale gray*, uncorrected; *dark gray*, corrected); underneath, the resulting structure after exposure is shown. The electron microscope pictures show a comparison of an uncorrected mask with a segment-mode-corrected mask [19]

Fig. 19.2). Although simulation models are used to correct the mask data prior to the first exposure, iteration steps to obtain the correct mask exposure are normally necessary.

All these methods have been introduced to decrease k_1 further by about a factor of 2. In Fig. 19.1, the implementation of these methods can be seen from the divergence of the curve with square points and the curve with circular points. In practice, the value of k_1 can currently reach values as small as 0.3. On the other hand, these mask-correcting measures have significantly aggravated the problem of reticle fabrication: controlled chromium and glass etching are necessary, and smaller line widths are required for the optical

proximity corrections. This has raised the cost of the masks, and often the time required for the generation a mask set is of the order of weeks.

19.3 Next-Generation Lithographies

The end of optical lithography for structure definition has been expected for some time because a steady change of the wavelength and such a strong improvement of the k factors were not anticipated. All alternative lithography methods, *next-generation lithographies*, are targeted at a significant decrease of the wavelength. For photons, hard and soft x-rays can be considered, but particles, such as ion and electron beams, are also suitable for lithography. All types of lithography have been intensively studied during the last 20 years, and the options for mass production have been narrowed down to the method of extreme UV lithography, also referred to as soft x-ray lithography. Electron beam lithography, one of the oldest and best-known lithographic techniques for mask production, is about to be deployed for rapid prototyping or low-volume production to avoid huge mask costs. The main hurdles for other lithographic techniques are direct (proximity) printing, which means that 1:1 masks ($1\times$) are needed, and the Coulomb interaction in charged-particle lithographies, which leads to a resolution deterioration when high particle densities or currents are used and limits the throughput.

X-ray lithography with photon wavelengths of about 1 nm has been under development for about 20 years now. There are no suitable lenses for this wavelength regime available, so 1:1 proximity printing is necessary. The costs, production time and short degradation time of the $1\times$ mask are the main roadblocks for this technique. The difficulties with the light source – synchrotron radiation is normally used – should also be kept clearly in mind.

Charged-particle projection lithographies have also been developed in recent decades with promising results but have been skipped by most semiconductor companies and suppliers in favor of EUV lithography. One major difficulty is the Coulomb interaction, which causes the particles to repel each other, resulting in a resolution deterioration by beam broadening (the Loeffler effect, perpendicular to the direction of propagation) and energy broadening (the Boersch effect, in the direction of propagation), causing a deterioration in the performance of the lenses. For both electron and ion beam lithography systems, stencil masks have been considered: the particles either pass through free spaces (holes) in the mask or are stopped by an absorber material. This causes the doughnut problem, which prohibits embedded absorber spaces and can be overcome by the use of two masks or the use of struts in the mask and therefore requires a scanning of the mask. Another variant is the use of absorber structures on a thin membrane which the particles can pass through and which, however, results in a worse resolution owing to scattering.

Ion beam projection lithography was the subject of a European project between IMS Vienna and Infineon Technologies that delivered good results

but which was stopped in 2001. Electron beam projection with scattering masks (SCALPEL, scattering with angular limitation projection electron beam lithography) was developed and investigated at Lucent Technologies' Bell Laboratories [5] but was also stopped in 2001. Finally, the electron projection system PREVAIL (projection electron variable-axis immersion lens) of IBM and the LEEPL (low-energy electron projection lithography) project of a Japanese consortium are under development. These technologies are described briefly in the next section.

EUV technology will be the dominant future production technology because Intel and almost all American and European semiconductor companies have decided to support the research activities in this sector and have joined a consortium called EUV LLC (EUV Limited Liability Corporation) to realize this next-generation lithography. Although tremendous efforts have been made, the launch of this technique is likely to be set back. Figure 19.3 shows a schematic illustration of an EUV stepper. One possible light source is a laser plasma source with a wavelength of 13.4 nm. The light is produced by laser illumination of xenon atoms, where various xenon sources are possible, e.g. Xe droplet sources or filament jets. The conversion efficiency of irradiation laser power to output EUV power is about 1%. Transmission optics are not possible, so reflection optics like those in radio telescopes are used. One lens consists of approximately ten double layers of molybdenum and silicon with a thicknesses of 6.9 nm per double layer. The thickness homogeneity must be very precise over the lens area, and the interface roughness must be tightly controlled. The reflectivity is limited mainly by the interdiffusion of silicon and molybdenum atoms and surface oxidation. The surface oxidation can be dealt with by an optimized capping layer (e.g. ruthenium) and the interdiffusion can be reduced by additional diffusion barriers between the layers. A measured reflectivity of around 70% has been achieved, compared with the ideal reflectivity of 74%. The use of eight lenses in the optical system results in a photon yield of 5.8% at the wafer.

The mask works in principle like the optical lenses: a structured Mo/Si multilayer either reflects locally or serves as an absorber. The mask is built up on a silicon wafer by conventional semiconductor techniques. The challenges concerning the reticle include heat dissipation and the resulting image distortion due to deformation of the mask. The required flatness of the mask and of the wafer necessitate the use of an electrostatic wafer chuck to achieve a flatness of better than 200 nm.

Not only the multiplicity of the new elements but also the tough specifications of the individual components require substantial development effort. The ongoing work to increase the efficiency of the light source could cause a delay in the introduction of this technique into production. As already mentioned, this technique will be indispensable for high-volume production, but might also be appropriate for low-volume production and prototyping.

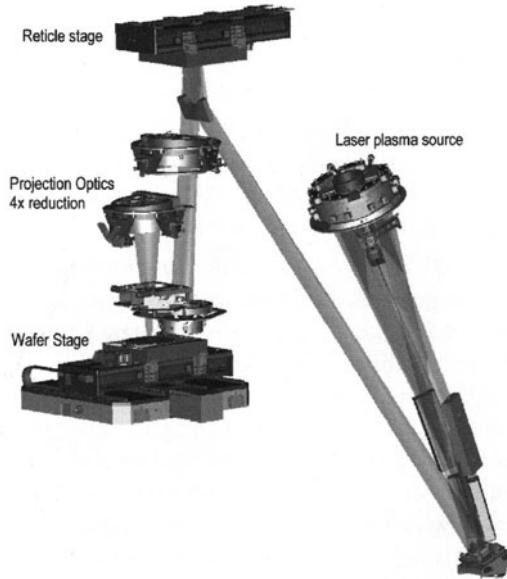


Fig. 19.3. Schematic illustration of the prototype of the EUV exposure tool of EUV LLC. There are two different modules, the main module containing the reticle and wafer stage and the reduction optics, and a separate illumination module housing the light source

19.4 Electron Beam Lithography

Electron beam lithography is the oldest lithographic method with the exception of optical lithography. In 1973, the first electron beam lithography system, based on a scanning electron microscope, was used to fabricate transistors with a $1\text{ }\mu\text{m}$ gate length [2]. Just four years later, ETEC introduced the first commercial mask writer, MEBES. Electron beam lithography has been used since then as a tool for the production of state-of-the-art photomasks. In particular, the decrease of the k factors by enhancement techniques, as mentioned Sect. 19.2 on “optical lithography”, demand high resolution “sub”-lithographic features with excellent line edge roughness and feature size constancy.

There are several possibilities for the generation of electrons. For lithographic purposes, thermal field emission (TFE) cathodes are usually used, where the electrons are extracted from a tip by means of a strong electric field and emission is enhanced by heating the tip. In contrast to cold field emission, this electron generation method yields a higher current that is stable over long periods. The energy spread for TFE is marginally larger than that for cold field emission.

The wavelength of electrons is calculated using the de Broglie formula

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2mE}}, \quad (19.3)$$

and is in the picometer range for electron energies of some tens of keV. However, the spot size of a Gaussian electron beam is limited by the aberrations of the electron lenses. In current scanning electron microscopes, electron spot sizes of around 1 nm can be achieved. Electrons are deflected by the Lorentz force $\vec{F} = e(\vec{E} + \vec{v} \times \vec{B})$, so electron beam imaging can be realized by electrostatic or electromagnetic lenses. Like their optical counterparts, electromagnetic lenses show errors such as spherical aberration, astigmatism, coma, chromatic aberration and diffraction. Some of these errors can be corrected, such as isotropic astigmatism and distortion; others can be minimized by measures that limit either the resolution or the throughput. This mutual exclusion is a crucial restriction on all charged-particle lithographies. Minimization of the spherical aberration by a reduction of the aperture angle (smaller apertures) down to the diffraction limit reduces the number of electrons and restricts throughput. Larger apertures enable larger throughput at the cost of resolution. In addition, when very high particle densities are employed, the Coulomb interaction between the electrons becomes relevant, causing beam broadening (transverse trajectory displacement, Loeffler effect) and energy broadening (longitudinal displacement, Boersch effect), causing an aggravation of the chromatic error.

Another important component of an electron beam writer is the deflection element. Deflection must be fast and accurate, so electrostatic deflectors should be used. Magnetic deflectors suffer from hysteresis and are limited in their speed. This is the reason why adapted electron microscopes, which are usually equipped with magnetic deflection systems, are limited in their writing speed to some tens of MHz and are therefore only suitable for research and development purposes. Further, a beam deflection element before the final lens causes additional lens errors, since the electrons are shifted away from the optical axis. These additional aberrations have to be compensated. Alternatively, an in-lens deflection system can be used. Finally, a fast beam blanker is necessary to ensure fast on-off switching of the beam.

There are two main exposure strategies (Fig. 19.4): pattern exposure with either a Gaussian or a shaped beam. In the Gaussian-beam case a focused electron beam writes the pattern pixel by pixel sequentially, like a paint-brush. The resolution in this case is defined by the beam diameter. These methods are substantially slower than the shaped-beam methods, but offer the best resolution. They are primarily used in research and development for the creation of single nanoscale device elements. Resolutions of several nm in resists have been shown and silicon transistors with a gate length of 8 nm have been fabricated using the Gaussian-beam technique. A further advantage of this resist-based technique is the possibility of integrating it into a production facility.

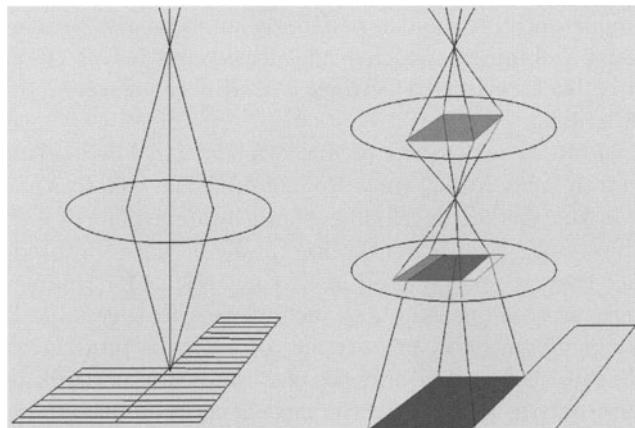


Fig. 19.4. Principle of Gaussian-beam and variable-shaped-beam electron beam lithography. In the Gaussian-beam case (*left*), a focused beam writes the structure line by line sequentially. In the shaped-beam case (*right*), a beam “stencil” is formed by projection of two apertures onto the substrate

Shaped-beam methods are the state of the art for mask-making purposes. In this case an expanded beam with a homogeneous current density is formed by two square apertures or a patterned aperture plate. This geometric structure is projected, scaled down, onto the wafer. In this manner, larger structures are printed onto the resist, speeding up the exposure time. The definition of resolution in this case is given not by the beam size but by the edge acuity, which is dominated by the lens and aperture aberrations.

Another issue in electron beam lithography is the interaction of the electrons with the resist and the sample. Owing to scattering processes, the primary electrons penetrating the resist layer are diverted, leading to a broadening of the latent image in the resist. Additionally, owing to their large range, the primary electrons penetrate the substrate, generating secondary electrons that enter the resist and also contribute to the exposure. The forward scattering of the electrons in the resist is a few nanometers, whereas the area covered by backscattered secondary electrons is some tens of μm^2 , depending on the primary energy. This effect is called the proximity effect, not to be confused with the optical proximity effect mentioned above. This proximity effect can be taken into account during the exposure by assuming a certain distribution function of the scattered electrons and adapting the dose of the written structure to achieve a uniform electron yield over the structure. For this purpose the original structure is segmented into smaller elementary structures, whose doses are allocated by the proximity correction software.

At present the trend in mask-writing machines is towards higher energies. The advantage of this is the minimization of the chromatic aberration,

but, more importantly, a smaller scattering angle of the primary beam in the resist can be obtained, which leads directly to better resolution. The contribution of backscattered electrons is routinely corrected by proximity correction programs.

The demanding requirements of mask generation have led to great advances in electron beam lithography. Reduction of the k factors has prompted a great leap in the further development and improvement of electron beam lithography systems. Even direct-write methods have profited from this progress, since they are partly used in writing reticles for leading-edge technologies. This has also pushed these technologies further into the realm of possibility for deployment in prototyping or even for production. This becomes manifest in the fact that only the electron beam methods are still alive as possible alternatives to EUV for the next-generation lithographies. In this case, the projection methods PREVAIL and LEEPL are possible options. PREVAIL is a further development of a shaped-beam system, projecting not the pattern of a predefined aperture plate but parts of a stencil mask. The deflection system works inside an electron immersion lens, which is an IBM invention called VAIL (variable-axis immersion lens). For the projection of a part of the $4\times$ reticle, three lenses of this type are used: two for the deflection of the electrons through the desired mask sector and one for the placement on the wafer (Fig. 19.5). The calculated throughput for this method is 20 wafers/hour (200 mm) and is therefore very promising. An IBM cooperation with Nikon has demonstrated an R&D exposure tool for the 70 nm node.

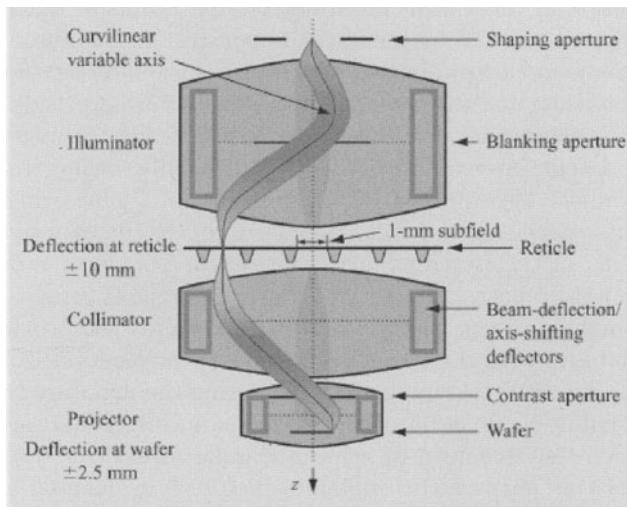


Fig. 19.5. Principle of PREVAIL [21]

The alternative, proposed by a Japanese consortium, is the LEEPL technique [16]. In this case, low-energy electrons of energy 2 keV are used for proximity printing of a 1× stencil mask onto a substrate (Fig. 19.6). The members of the consortium benefit from knowledge obtained from extensive R&D experiments in proximity x-ray printing. The electrons are produced in miniaturized electron beam columns. These are parallelized in order to increase throughput. The low-energy electrons minimize the proximity effect, but increased forward scattering limits the resolution. This problem can be attenuated by using thinner resist films. The manufacturers claim that there is no space charge effect which could limit throughput. One of the printing problems is that a 1× mask similar to that used in a former x-ray proximity lithography technique is needed. This was one reason why that particular x-ray lithography was halted, because the production of the stencil masks was too demanding and costly. In the meantime, however, Gaussian beam electron lithography has improved so that it no longer seems impossible to produce masks like this.

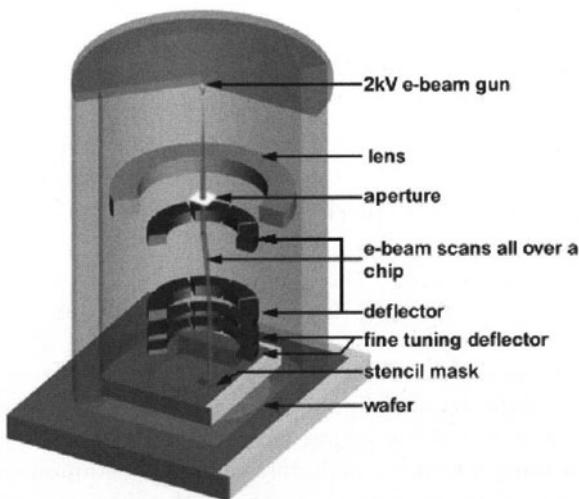


Fig. 19.6. Principle of LEEPL [20]

There is also one electron beam lithography concept for the more distant future which does without masks. This type of lithography is called ML2 (**maskless lithography**). Leica together with IMS Vienna has announced a collaboration to study the feasibility of a programmable aperture plate. The “programmed” chip image is then scaled down 200 times by reduction projection optics [15]. A maskless lithography method would solve the problems of rapidly increasing mask costs and shorten the time for prototyping new components.

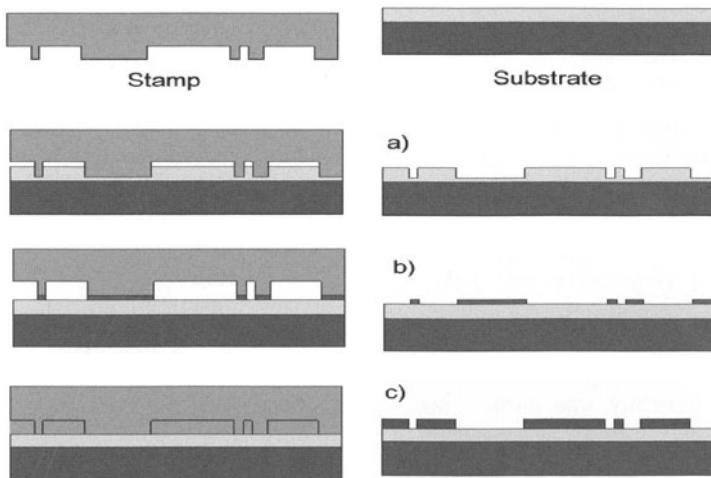


Fig. 19.7. Schematic illustration of three imprint methods. (a) Nanoimprint. A mask is stamped onto a polymer-coated substrate at elevated temperature and pressure, forming an image of the stamp in the resist. (b) Inking. An ink is transferred from the stamp to the substrate. (c) Step and flash imprint. A mask is brought into contact with a sample coated with a liquid photopolymer, which is exposed during the stamping process with a UV flash exposure. Thus the solidified polymer remains on the substrate

19.5 Nanoimprint Methods

Nanoimprint lithography spans the technology where a pattern is stamped from a mask directly onto the substrate. Three methods can be distinguished (Fig. 19.7). The first method, called Imprint [23], is similar to the production technique for CDs and DVDs, in which a master mold, the mask, is nanofabricated by electron beam techniques and etching and then stamped onto a polymer resist layer. For this purpose, the polymer layer is heated above the glass transition temperature immediately before the stamping process, and the stamping process takes place at pressures of several atmospheres. After the stamp and the substrate are brought into contact, the resist layer is cooled down below the glass transition temperature and the stamp is removed. Further processing for pattern transfer can be applied afterwards. Sub-10 nm features have been produced with this technique on a laboratory scale [24]. A challenge with this technique is that the polymer displaced during the stamping process needs a place to flow into, either in the stamp or in the sample. This complicates the layout when one is producing dense structures. Further issues are an antisticking layer between the mold and the polymer, and the thermal gradient during the printing process, which leads to distortions and alignment errors. Two variants of this technique, which avoid these problems are inking techniques [17] and “step and flash imprint lithography”

(SFIL) [18]. Inking is similar to the stamping process: an ink, normally a self-assembled monolayer, that has been selectively applied to the stamp is transferred to the substrate. Since functionalization of either the stamp or the sample is necessary, and the transferred layer is very thin, complicating the succeeding transfer steps, this technique is not likely to be used for production in the near future. The SFIL process utilizes an additional etching barrier between the resist and the transparent stamp (e.g. quartz), which consists primarily of a photopolymer. The stamping process takes place at low pressure and ambient temperature, and polymerization of the etching barrier is performed by UV flood exposure through the quartz blank. This technique circumvents the temperature problem and the necessity for disposing of the displaced polymer.

All nanoimprint lithography methods suffer from the fact that a $1\times$ mask is necessary, and is also perishable. In addition, a reliable alignment method with a 10 nm overlay capability has not been realized. Further, the mask errors or distortions are transferred directly to the substrate and there is no possibility of intermediate correction, which is feasible in PREVAIL and LEEPL, where known errors in the mask layout can be corrected in the projection process. Nevertheless, the imprint principle can provide the highest throughput conceivable and is, therefore, the subject of intensive research activities. Through advancements in the fabrication accuracy of $1\times$ masks, which are also necessary for LEEPL, these imprint techniques could become more viable for use in production in the future.

19.6 Proximal-Probe Lithography

Proximal-probe lithographies, or scanning probe methods, offer the ultimate resolution of all lithography techniques. Scanning probes in scanning tunneling, scanning force or scanning near-field microscopes interact with a very small area of the sample surface, which permits the manipulation of even single atoms. The most impressive results are those of Eigler and Schweizer, who arranged single atoms of Xe on a Ni surface with an ultra low-temperature STM (Fig. 19.8) [3].

There are many possibilities for proximal-probe lithographies, working either in bottom-up mode, i.e. building up a structure from single atoms, molecules or clusters, or in top-down mode, where a layer is structured laterally by the scanning probe. Apart from the manipulation of single atoms, the bottom-up methods include the deposition of structures by the tip material [6] or by the decomposition of adsorbed gas molecules (a CVD like method) [7]. The structuring of a resist layer [8] and the oxidation of a previously deposited layer, e.g. carbon [9] (which acts as a self-developing positive resist) or Ti (for the direct fabrication of a single-electron transistor) [14], belong to the top-down methods. Recently, even combinations of scanning probe methods with stamping or inking technologies have been demonstrated. By structuring a

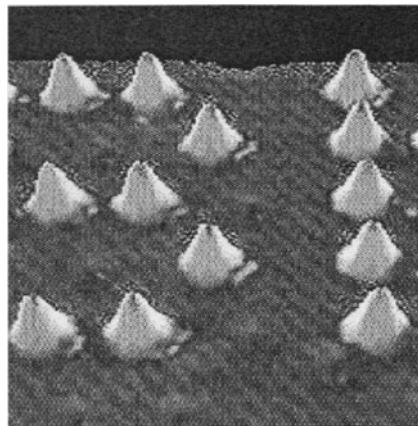


Fig. 19.8. Atom lithography with xenon on nickel by IBM [22]

blunt STM tip, a predefined structure can be stamped many times onto a substrate [10], for example by the oxidation of a carbon layer. The combination of inking technology with the SFM, scanning force microscopy is called Dip Pen Lithography [11], where the tip serves as a source for single molecules that are transferred to the sample across a water meniscus, which shapes itself between the tip and the sample under ambient conditions.

All proximal-probe lithographies have been used to produce remarkable devices with outstanding results on the laboratory scale. Since these techniques are about a factor of 1000 slower than direct-write Gaussian electron beam lithography, their application in development is at present inconceivable, not to mention their use in production. There are approaches that use parallel cantilevers and tips to parallelize the processes, but the fabrication and control of some thousand cantilevers at the same time is a big challenge, on the one hand with regard to the fabrication difficulties, and on the other hand in organizing the huge data flow which would be necessary. Further, redundancy has also not been considered yet, but this has to be taken into account if one of the cantilevers fails. Nevertheless, parallel cantilevers have been fabricated for the purpose of testing lithographic processes at Quate's group at Stanford University [13]. The most remarkable application of parallel cantilevers is the milliped project at IBM Rüschlikon [12]: an array of 1024 cantilevers, which can be individually controlled simultaneously in the vertical and lateral directions, are used to store single bits by indentation in a polymer layer. Readout and erasing of bits are performed with the same cantilevers. With this instrument, a data density of 31 gigabits per square centimeter has been demonstrated, which means a distance of around 60 nm between two bits.

19.7 Conclusion

Lithography is the key technology for the production of silicon devices. Optical lithography has dominated production since the beginning of production, in spite of all of the predictions of its demise in recent decades. EUV will be the successor to optical technology. The immense expense of this technology, with regard to both the provision of this technique and the mask costs, makes it necessary to study other fabrication techniques. The current knowledge in the field of electron beam lithography predestines this method for rapid prototyping and low-volume production at its present state of maturity. Advanced electron beam methods for mass production are currently under intense study. They could be an alternative to EUV lithography, although throughput is limited in charged-particle optics owing to the Coulomb interaction. Methods requiring $1\times$ masks are very challenging; x-ray lithography has been stopped in former times owing to hurdles in the fabrication of these masks. Nevertheless, imprint methods with $1\times$ masks are attractive because of their high throughput potential. Last but not least, proximal-probe lithographies are essential for research fabrication of nanodevices, but their state of maturity at this time excludes employment in the near future.

References

1. ITRS, *International Technology Roadmap for Semiconductors*, update (2002), <http://public.itrs.net>
2. F. Fang, M. Hatzakis, C.H. Ting: J. Vac. Sci. Technol. **10(6)**, 1082 (1973)
3. D.M. Eigler, E.K. Schweizer: Nature **344**, 524 (1990)
4. C.R. Marrian (ed.): *Technology of Proximal Probe Lithography* (SPIE Press, Bellingham 1993)
5. S. Berger, C. Biddick, M. Blakey, K. Bolan, S. Bowler, K. Brady, R. Camarda, W. Connelly, R. Farrow, J. Felker, L. Fetter, L. Harriott, H. Huggins, J. Kraus, A. Liddle, M. Mkrtchyan, A. Novembre, M. Peabody, T. Russell, W. Simpson, R. Tarascon, H. Wade, W. Waskiewicz, P. Watson: Proc. SPIE **2322**, 434 (1994)
6. H.J. Mamin, P.H. Guethner, D. Rugar: Phys. Rev. Lett. **65(19)**, 2418 (1990)
7. E.E. Ehrichs, W.F. Smith, A.L. de Lozanne: Ultramicroscopy **42-44**, 1438 (1992)
8. M.A. McCord, R.F.W. Pease: J. Vac. Sci. Technol. B **619**, 293 (1988)
9. T. Mühl, H. Brückl, G. Weise, G. Reiss: J. Appl. Phys. **82(10)**, 5255 (1997)
10. T. Mühl, J. Kretz, I. Mönch, C.M. Schneider: Appl. Phys. Lett. **76(6)**, 786 (2000)
11. R.D. Piner, J. Zhu, F. Xu, S. Hong, C.A. Mirkin: Science **283**, 661 (1999)
12. P. Vettiger, M. Despont, U. Drechsler, U. Durig, W. Haberle, M.I. Lutwyche, H.E. Rothuizen, R. Stutz, R. Widmer, G.K. Binnig: IBM J. Res. Devel. **44** (3), 323 (2000)
13. S.C. Minne, P. Flueckiger, H.T. Soh, C.F. Quate: J. Vac. Sci. Technol. B **13** (3), 1380 (1995)

14. K. Matsumoto, M. Ishii, K. Segawa: *J. Vac. Sci. Technol. B* **14** (2), 1331 (1996)
15. H. Loeschner, G. Stengl, H. Buschbeck, A. Chalupka, G. Lammer, E. Platzgummer, H. Vonach, P.W. de Jager, R. Kaesmaier, A. Ehrmann, S. Hirscher, A. Wolter, A. Dietzel, R. Berger, H. Grimm, B.D. Terris, W.H. Bruegger, D. Adam, M. Boehm, H. Eichhorn, R. Springer, J. Butschke, F. Letzkus, P. Ruchhoeft, J.C. Wolfe: *Proc. SPIE* **4688**, 595 (2002)
16. T. Utsumi: *Jpn. J. Appl. Phys.* **38**, 7046 (1999)
17. D. Wang, S.G. Thomas, K.L. Wang, Y. Xia, G.M. Whitesides: *Appl. Phys. Lett.* **70(12)**, 1593 (1997)
18. M. Colburn, S. Johnson, M. Stewart, S. Damle, T. Bailey, B. Choi, M. Wedlake, T. Michaelson, S.V. Sreenivasan, J. Ekerdt, C.G. Willson: *Proc. SPIE* **3676**, 379 (1999)
19. Courtesy of Infineon Technologies: A. Eckardt, L. Tarek
20. Reproduced from LEEPL website (www.leepl.com)
21. Reproduced from IBM Research website (www.research.ibm.com)
22. Reproduced from IBM Research website (www.almaden.ibm.com)
23. S.Y. Chou, P.R. Krauss, P.J. Renstrom: *Science* **272**, 85 (1996)
24. S.Y. Chou, P.R. Krauss, W. Zhang, L. Guo, L. Zhuang: *J. Vac. Sci. Technol. B* **15(6)**, 2897 (1997)

Further reading and overview articles:

25. P. Rai-Choudhury (ed.): *Handbook of Microlithography, Micromachining and Microfabrication* (SPIE Optical Engineering Press, Bellingham 1997)
26. D. Widmann, H. Mader, H. Friedrich: *Technology of Integrated Circuits*, Springer Series in Advanced Microelectronics vol. 2 (Springer, Berlin, Heidelberg 2000)

20 Silicon Sensors

E.F. Krimmel

20.1 Introduction

Sensors are used to monitor important parameters. Sensors are used in the areas of biochemistry, biophysics, medicine, genetic diagnostics, the diagnosis of asthma, pollution control (e.g. herbicides, fungicides and insecticides) and fireprevention, and to control process conditions, to ensure safety in manufacturing, to monitor hazardous radiation and even to monitor the aroma of olive oil. Sensors can be grouped into two types, namely “chemical” sensors and “physical” sensors. Silicon-based sensors have a particular attraction owing to the possibility of monolithic integration with microelectronic circuits and consequently the possibility of fabricating multifunctional devices, which may even be suitable for implantation into the human body. Because of its favorable chemical and physical properties, silicon nitride plays a fundamental role in both the fabrication of sensors and the use of sensors. We may say that silicon nitride is the heart of the device. For this reason, the discussion here will concentrate on the use of silicon nitride in sensors.

The first relevant work was made public by a patent in 1964 [1], followed immediately by some papers in 1965 and 1966 [2–4]. The real boom in the utilization of silicon nitride, which can be represented as Si_3N_4 , but perhaps better as Si_xN_y , started at the end of the 1970s and the beginning of the 1980s. Particularly intensive activities can be recognized in Japan, demonstrated by an enormous number of disclosures.

20.2 “Chemical” Sensors

Sensors which monitor ions usually belong to the class of FET-type devices, the ISFETs (ion-sensitive FETs). Their gates are of SiO_2 or Si_3N_4 . The gate serves as the sensing element. However, the conventional gate metal may be replaced by an electrolyte, which must be free of interfering cations such as Na^+ , K^+ and Ca^{2+} , when pH measurements, for example, are to be performed (see Fig. 20.1). A specific species-selective membrane covering the gate serves to identify ions, atoms and molecules in gaseous or liquid environments. The species to be monitored are selectively adsorbed on the gate dielectric. The electrochemical potential, i.e. the gate potential at the interface between the

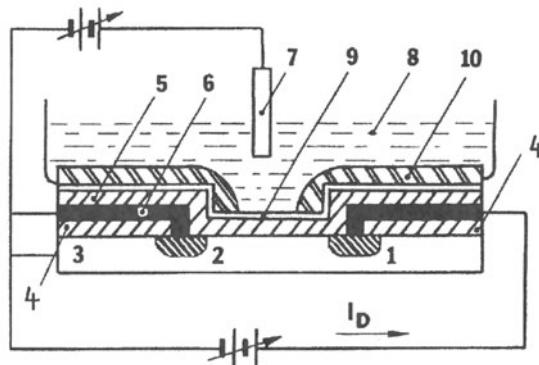


Fig. 20.1. Schematic illustration of a Si_3N_4 ISFET pH sensor. (1) drain, (2) source, (3) Si substrate, (4) insulation layer, (5) SiO_2 inner gate insulation layer, (6) metallic contacts, (7) reference electrode, (8) aqueous solution of interest, (9) thin Si_3N_4 layer on the gate insulation, (10) encapsulation layer

medium and the gate dielectric, adjusts itself according to the ion activity of the solution by polarization, depending on the adsorbed species. However, such sensors may have a limited lifetime owing to irreversible poisoning. Furthermore, significant hysteresis effects have been ascribed to changes in the depth of ionic penetration in the Si_3N_4 layer [5]. Some films, for instance, have shown hysteresis of 0.02 to 0.06 pH units. There is another, quite elementary-looking (it is not!) problem: how to store the sensors when they are not in use? It seems, for example, that storage in an aqueous solution of sodium silicate is reasonable.

The ISFET structure may consist of a glass substrate, an Al electrode, a SiO_2 and/or a Si_3N_4 layer, and an amorphous Si layer deposited by PCVD. ISFETs coated with $\text{Si}_3\text{N}_4/\text{SiO}_2$ gate dielectrics show good ion selectivity in electrolytic solutions. ISFETs on Si substrates with V-shaped grooves are also known. Gas sensors prepared on heavily B-doped Si substrates are coated with a Si_3N_4 layer and overcoated with a sensor film. Gas sensors may also contain Schottky barrier diodes.

Insulating Si_3N_4 layers in ISFETs may be doped, for instance, with Al and K by ion implantation and annealed (see Chap. 11) to prepare alkali-ion-selective membranes. The sensing layers of Na^+ -sensitive ISFETs are fabricated by implanting Na^+ into the surface of an oxidized silicon nitride layer through an Al buffer layer. The buffer layer reduces the radiation damage to the gate insulator. The sensitivity is nearly Nernstian, i.e. it has a slope to 58.6 mV/pH unit at 22°C [22], and independent of the pH value within the range from 7 to 10. The period of long-term stability has been determined as 1300 h [6].

The gate insulators of K^+ sensors have a 60 nm thick silicon nitride top layer covered with a poly(vinyl chloride) membrane. The poly(vinyl chloride)

layer is loaded with liquid ion exchangers. However, interference from species such as CO_2 , benzoic acid and benzoates, which permeate the membranes, is observed [7].

Top layers loaded with liquid ion exchangers have also been reported in relation to Ca^{2+} sensors [8]. Ag^+ ion-selective membranes have been made from thin films of $\text{AgCl}-\text{AgBr}$ mixtures [9].

F^- sensors with LaF_3 layers have a measuring range from 10^{-1} to 10^{-5} M [10]. Cl^- and Br^- ion-selective layers in Cl^- and Br^- sensors consist of thin films of an $\text{AgCl}-\text{AgBr}$ mixture [9]. Even ion-sensitive membranes consisting of a polymer-containing quaternary ammonium chloride and dodecylbenzenesulfonic acid for use in a bioambient have been claimed. No interference of catalase, albumins, etc. was observed [11].

The selectivity mechanism of hydrogen-ion-sensitive ISFETs investigated, for instance, on Si/SiO_2 /electrolyte and $\text{Si}/\text{SiO}_2/\text{Si}_3\text{N}_4$ /electrolyte systems shows a linear dependence of the flat-band voltage when the pH varies over a range from 2 to 10, and the sensitivity is 30–40 mV/pH and 45–55 mV/pH, respectively. The complex ion sensitivity mechanism is attributed mainly to ion-exchange processes at the insulator/electrolyte interface [21]. The pH sensitivity of sensors with Si_3N_4 layers is attributed to surface properties, i.e. properties of surface sites of the films, involving NH groups and SiH groups present in the film. The active sites on the Si_3N_4 surface include OH and NH_2 groups [23].

Hydrogen gas sensors and pH sensors of the tunnel MIS diode type have a Si_3N_4 tunnel layer and may consist of a $\text{Pd}/\text{Si}_3\text{N}_4/\text{Si}$ structure [12]. Both the high-impedance OFF state and the low-impedance ON state depend on the hydrogen concentration. The capacitance of such MIS diodes, which is proportional to the concentration (ppm) of hydrogen, can be used to control hydrogen.

A further development is represented by CMOS integrated chemical sensors containing pH-sensitive ISFETs with B-doped Si_3N_4 layers in order to tailor the ion sensitivity or to obtain differential sensing [15, 16]. Integrated chemical sensors to monitor pH, Na^+ , K^+ and Cl^- , composed of four Si_3N_4 gate ISFETs, a Clark-type gas sensor, and a Severinghaus-type gas sensor using pH ISFETs, have been reported. The ISFET sensors have a sensitivity of 50 mV/decade and a linear range between 10^{-4} and 5×10^{-1} mol/L. The Clark-type gas sensor has a sensitivity of 0.35 nA/mm Hg, and the Severinghaus-type gas sensor 42 mV/decade; their response times are 30 s and 1 min, respectively [24].

Other known sensors monitor the concentration of oxygen, the humidity, the concentration of CO_2 in the blood, and the concentration of gases such as B_2H_6 , SiH_4 , PH_3 and AsH_3 used in the electronics industry.

A theoretical “induced-dipole model” has been claimed to be confirmed by measurements of the flat-band voltage shift of $\text{Pd}/\text{Si}_3\text{N}_4/\text{SiO}_2/\text{Si}$ MIS structures [13]. The mechanism of H^+ and OH^- ion-sensitive ISFETs has been discussed in [14] and a relevant “two-side theory” has been presented in [17].

20.2.1 Biosensors

Biosensors are sensitive to heat and chemical stress and hence cannot always be used directly to monitor processes such as fermentation. The problem is that the enzymes used have only a short lifetime and tolerate only low temperatures, say 37°C. For example, systems to control fermentation by determining the glucose and ethanol concentrations using a dehydrogenase that depends on NAD⁺ (nicotinamide adenine dinucleotide) consist of a detection part, i.e. an immobilized enzyme and a sensing head for NADH (+H⁺), and a quite complex transport system for the substance to be analyzed. The enzymes, provided from a solution, are stored in a membrane, which is provided with active centers to bind the enzymes in order to immobilize them. Sensors used in biochemistry utilize membranes containing immobilized enzymes, for example to decompose complex molecules such as glucose to be monitored as H₂O₂.

The concentration of complex molecules is related to the concentration of the decomposed simpler species. For instance, ISFETs may be coated with a poly(vinyl butyral) membrane to measure adenosine triphosphate (ATP). Other sensors are used to monitor threonine and urea in the blood plasma. The implantation of biosensors for long-term control of physiological parameters is of great medical interest and therefore also of industrial interest.

Piezoresistive or elastic cantilevers with a thickness of the order of micrometers can be utilized for biosensing applications by detecting changes in surface stress due to binding and hybridization of biomolecules.

20.3 “Physical” Sensors

The parameters to be determined with physical sensors are quite heterogeneous, ranging from pressure to radiation. Consequently, a large variety of different types of sensors is utilized.

For instance, sensors that respond to vibration or to pressure are usually based on the principle of bars or membranes. The membrane may be prepared by depositing a Si₃N₄ film on the front surface of a Si substrate and selectively etching a window into the substrate from the rear to expose the film, thus forming a membrane (see e.g. [18]). Pressure transducers may contain an etch-resistant Si₃N₄ layer on a piezoresistive Si substrate [19].

Airflow microsensors may also be based on thin-film diaphragm pressure sensors, or may be based on the measurement of the temperature-sensitive electrical resistance. Temperature and IR sensors may be prepared on heavily B-doped Si substrates coated with a Si₃N₄ layer and overcoated with a sensor film, or may consist of thin Si resistor tracks. Such sensors are applied, for example, to record the temperature of heated specimens during MBE.

Radiation-sensing FETs (RADFETs) may consist of a Si substrate coated with a thick Si₃N₄ layer and a high-quality thermal SiO₂ layer. Trapping layers at the Si₃N₄/SiO₂ interface store the radiation-generated positive charge.

Such RADFETs are stable and can have a sensitivity of $86 \mu\text{V}/\text{rad}$ dose at room temperature [20]. Detectors for γ -rays, γ/n radiation and α particles, for example may be prepared using complex multilayer systems including Ta and Al. Hall sensors fabricated on GaAs:Cr substrates are passivated by Si_3N_4 layers.

20.4 New Ideas and Developments

The devices discussed in the foregoing paragraphs are not the end of the development of sensors. Silicon offered us natural SiO_2 , one of the basic reasons for its success from the very beginning. Now we are making a great leap forward. Silicon offers us also the porous state of silicon, which is easily accessible, e.g. by anisotropic electrochemical etching of (110) silicon wafers. Hence it is no miracle that part of today's human creativity has been focused on porous silicon as an economic base material since this porous state became known; see Chap. 8 and [25]. A basic target for the future is to drastically increase the active sensing surface to hundreds of square meters per cubic centimeter in order to obtain increased sensitivity and to reduce the volume of the specimen at the same time. Which of all the relevant efforts in basic research and development will finally win through and become an industrial product available for common use cannot be predicted. It is interesting to recognize simply that porous silicon seems to play a major role.

The surface of porous silicon, which is in general covered by Si:H_x species, oxidizes slowly even at low temperatures. The surface can be treated with molecules which hinder further degradation and serve at the same time as a functional sensing species. Selected nanoareas of porous silicon specimens, surface-passivated by hydrogen, must be prepared in this way so that they sense various parameters; such an approach is favorable for monolithic integration. The activation of such areas is achieved by modifying their properties by chemisorption of the appropriate molecules, which may be of biological origin or organic, for example unsaturated hydrocarbons. Gaseous- or liquid-phase sensors sensitive to alcohol, acids and aliphatic n-alcohols, for instance, are of interest. Such sensors work on the basis of FETs, where the conductivity of the channel in the single-crystal silicon under the porous silicon is influenced by the adsorbed molecules.

Another way to obtain signals is to utilize the photoluminescence of porous silicon, which depends on the type and concentration of the adsorbed species. However, it seems that not all aspects of this effect are fully understood yet. Birefringence due to interaction between Si nanocrystals and pores may be observed and can be modified by molecules inside the pores, which provides a tool for sensing atoms or molecules. Models to understand this effect have already been proposed.

Porous silicon will surely not be the only object of study in current and future research, development and production in the field of sensors. There

are also other important techniques and materials. However, porous silicon has become particularly attractive and deserves to be so.

References

1. Telegraph Condenser Co., Ltd., UK Patent 947 271 (1964); *Chem. Abstr.* **60**, No. 11490 (1964)
2. H.F. Sterling, R.C.G. Swann: *Sol. State Electron.* **8**, 653 (1965)
3. V.Y. Doo: *IEEE Trans. Electron Devices* **ED-13**, 561 (1966)
4. V.Y. Doo, D.R. Nichols, G.A. Silvey: *J. Electrochem. Soc.* **113**, 1279 (1966)
5. I.R. Lauks, J.N. Zemel: *IEEE Trans. Electron Devices* **ED-26**, 1959 (1979)
6. T. Ito, H. Inagaki, I. Igarashi: *IEEE Trans. Electron Devices* **ED-35**, 56 (1988)
7. X. Li, E.M.J. Verpoorte, D.J. Harrison: *Anal. Chem.* **60**, 493 (1988)
8. S.D. Moss, C.C. Johnson, J. Janata: *IEEE Trans. Biomed. Eng.* **BME-25**, 49 (1978)
9. R.P. Bak, Yu. Vlasov, D.E. Khekleman: *J. Appl. Chem. (USSR)* **52**, 2465 (1979)
10. J. Salardenne, J.P. Couput, C. Quet: *Fr. Demande 2 600 822* (1987); *Chem. Abstr.* **109**, No. 47518 (1988)
11. Hitachi Ltd., Kokai Tokkyo Koho 58-167 951 (1982/83); *Chem. Abstr.* **100**, No. 99 455 (1984)
12. K. Murakami, T. Yamamoto, K. Doshita: *Proc. Electrochem. Soc.* **87-9**, 81 (1987)
13. A. Nannini, P.E. Bagnoli, A. Diligenti, V. Ciuti: *Solid State Electron* **28**, 867 (1985)
14. Yu.G. Vlasov, A.V. Bratov, Yu.A. Tarantov, M.P. Sidorova, I.B. Dimitrieva: *J. Appl. Chem. (USSR)* **61**, 2400 (1988)
15. T. Maruizumi, K. Tsukada, H. Miyagi: *Japan. Kokai Tokkyo Koho 61-181 954* (1986); *Chem. Abstr.* **107**, No. 32273 (1987)
16. H.S. Wong, M.H. White: *IEEE Trans. Electron Devices* **36**, 479 (1989); H.S. Wong, Y. Hu, M.H. White: *J. Electrochem. Soc.* **136**, 2963 (1989)
17. D.L. Harame, L.J. Bousse, J.D. Shott, J.D. Meindl: *IEEE Trans. Electron Devices* **ED-34**, 1700 (1987)
18. Rohm Co., Ltd., Japan. Kokai Tokkyo Koho 60-84 821 (1985); *Chem. Abstr.* **103**, No. 187 971 (1985)
19. P.E. Stevenson: US Patent 4721938 (1988); *Chem. Abstr.* **108**, No. 159 931 (1988)
20. R.C. Hughes, W.R. Dawes Jr., W.J. Meyer, S.W. Yoon: *J. Appl. Phys.* **65**, 1972 (1989)
21. Yu.G. Vlasov, A.V. Bratov, V.P. Letavin: *Anal. Chem. Symp. Series* **8**, 387 (1981)
22. A. Sibbald: *IEE Proc., Solid State Electron Devices* **130**, No. 5, 233 (1983)
23. N. Jaffrezic-Renault, A. De, P. Clechet, A. Maaref: *Colloids Surf.* **36**, 59 (1989)
24. K. Tsukuda, Y. Miyahara, Y. Shibata, H. Miyagi: *Sens. Actuators B* **2**, 291 (1990)
25. E-MRS 2003 Spring Meeting, June 10–13, 2003, Strasbourg, France, Symposium N, New Materials and Technologies. In: *Sensor Applications*, Organizers M. Koudelka-Hep, P. Siciliano

Part VIII

Supplementing Silicon: the Compound Semiconductors

21 Supplementing Silicon: the Compound Semiconductors

M. Jurisch, H. Jacob, T. Flade

21.1 Introduction

For a long time, the compound semiconductors, especially gallium arsenide, were believed to be the basic materials for a new generation of solid-state electronics which would replace silicon after the latter had reached its expected physical limits. This expectation turned out to be wrong. GaAs and other compound semiconductors did not replace silicon, but have found commercial application side by side with silicon, and have reached maturity and mass-production level for certain microelectronic and optoelectronic devices during the last decade of the last century. Therefore, compound semiconductors are no longer regarded as competitors to the elemental semiconductor silicon (and germanium), but as a necessary supplement, owing to some unique physical properties compared with silicon, although their production scale is and will be considerably less than that of Si. As the consumers of devices do not care about the technology utilized in the device, there are and will be overlapping fields of potential applications where devices based on silicon and on compound semiconductors will compete on the basis of performance, compatibility, price per die, etc. This has led in the past and will lead in the future to shifts of market shares between them. A convincing present-day example is the development of germanium-alloyed silicon, $\text{Si}_{1-x}\text{Ge}_x$, which is replacing the compound semiconductor GaAs in some areas of high-frequency microelectronics.

With respect to the position of the constituents in the periodic table of elements, the inorganic compound semiconductors relevant to applications can be classified into II–VI, III–V and IV–IV compounds, representatives of which are ZnSe, GaAs and SiC. Among them, the III–V compounds, particularly GaAs and GaN, are the most widely used compound semiconductors for the production of microelectronic and optoelectronic devices at present.

Owing to its superior role, GaAs has been chosen here as a prototype of compound semiconductors, and will be considered in greater detail in the following sections.

21.2 The Hard Way to the Successful Product

Despite some earlier “more stochastic” publications about III–V compounds and their properties, including semiconductivity [1,2], H. Welker is commonly regarded as the “father of the III–Vs”. From 1951 to 1953 he thought about gallium arsenide and its kin from the third and fifth groups of the periodic table of elements and recognized that the III–V compounds could be materials to compete with and/or to supplement silicon. A period of detailed investigations of the electronic and structural properties of GaAs followed, in which most of the leading semiconductor manufacturers were involved. Concepts of optoelectronic and microelectronic devices were developed and realized. The first lasing of a p/n junction was observed in 1961, and the semiconductor laser diode was born. The next decisive step was the development of the principle of heterojunctions on the basis of GaAs/GaAlAs structures by Krömer and Alferov. Combined with continuous progress in epitaxial growth, mainly by MBE and MOCVD, this resulted in the development of a new generation of devices for optoelectronic (double-heterostructure laser diodes) and microelectronic (HBTs, hetero bipolar transistor and HEMTs, high electron mobility transistor) applications, which are now produced on a large scale.

In parallel with this development, considerable efforts were undertaken to create the material basis, i.e. to synthesize GaAs and to grow single crystals for the manufacturing of substrates with the required properties.

The early and sometimes turbulent beginning of the pioneering work is recollected by one of us (H. Jacob), who was leadingly involved at Wacker in Burghausen, as follows:

At first sight, the synthesis of GaAs seemed to be as simple as the melting together of indium and antimony in the case of indium antimonide which was the first III–V compound used technologically for magnetic devices based on the Hall effect. But this was not so easy with gallium and arsenic. Put together the equivalent amounts of both components in a quartz tube, evacuate and seal it, and heat it up: it will explode. Gallium melts at about 30°C, and as the temperature rises it becomes covered by a thin layer of GaAs, inhibiting further reaction, whereas the arsenic evaporates, exerting a vapour pressure high enough to burst the ampoule at temperatures above 700°C. GaAs melts at about 1240°C, and it decomposes again if there is not an arsenic pressure over the melt. So more sophisticated arrangements had to be made. A carbon boat containing the gallium was put inside one end of a quartz tube, and a surplus of arsenic at the other end [see Fig. 21.6]. The quartz tube, evacuated and sealed, was heated at the gallium end a little above 1240°C, and the arsenic side was heated to a temperature of 600°C, where the arsenic pressure is about 0.1 MPa. Surprise, surprise: the ampoule exploded again.

Why? The arsenic commercially available at that time always contained oxide, which evaporates on heating and reacts with the carbon of the boat, forming carbon oxides. As three molecules of CO are formed from one molecule of As_2O_3 , too high a pressure was generated, which caused the ampoules to burst. So, until oxide-free, hyperpure arsenic was offered by manufacturers in sealed ampoules, one had to sublime the arsenic in a hydrogen atmosphere to make it oxide-free. However, whereas the ampoules no longer burst, the synthesized ingots after solidification either looked like a Swiss cheese or contained bubbles of gallium. Obviously, the vapour pressure of arsenic had to be adjusted more carefully, and it took a series of experiments with different arsenic temperatures and careful analysis of the GaAs ingots to determine the exact temperature at which the arsenic surplus had to be held. Sophisticated temperatures regulators and heat pipes were then in use.

Other problems arose. Carbon boats were too dirty, so boats made of pure quartz were used. But the GaAs melt tended to adhere to the quartz on crystallization. This was especially troublesome when growth of a monocrystal by slowly pulling the boat out of the high-temperature zone was intended. Coating of the boat with a silicon dioxide layer or sandblasting were for a long time the only remedies against that evil, until people found out about the reaction of gallium arsenide and quartz to form gaseous suboxides, the deposition of which in the colder zone of the tube must be avoided.

Once GaAs had been synthesized, single crystals had to be grown. The trouble with synthesis in a quartz boat led to many ideas to avoid crucible contact. The most obvious was to pull single crystals in the same way as silicon could be grown: by a crucible-free floating-zone technique. But, whereas silicon can be melted in a vacuum or in an inert atmosphere, the requirement for an arsenic atmosphere made the necessary equipment much more complicated for GaAs. From polycrystalline boat-grown GaAs ingots, square rods with a size of about 8×8 mm were cut. Such a rod was mounted in a quartz ampoule above a monocrystalline seed crystal. At the bottom of the ampoule, a surplus of arsenic was provided. The ampoule was evacuated, sealed and positioned in a furnace. By induction heating in a Keck apparatus (used at that time for silicon floating-zone growth in argon), the lower end of the polyrod was melted and brought into contact with the seed crystal. At the same time, the temperature of the furnace was controlled in such a way that selective arsenic evaporation from the molten zone was equalized by sublimed arsenic. The molten zone was moved upwards by lowering of the ampoule. Easier said than done! The density of molten GaAs is considerably higher than that of silicon, whereas the surface tension of the melt is lower. So

it was a feat to prevent the molten zone from dropping. Single (or often twinned) crystalline rods less than 100 mm in length and with maximum diameters of 8 mm were achieved at best, to the mockery of the silicon people, who were pulling crystals half a metre long and 20–30 mm in diameter at that time.

The Czochralski method was first successfully used to grow GaAs single crystals by Gremmelmaier [3]. Again, a hot-wall technology using a gas-tight quartz container was applied in order to prevent selective arsenic evaporation from the melt and condensation on colder parts. Seed and crucible translation and rotation were realized by a magnetic levitation system without any lead-throughs. Later on, this system was improved by Steinemann and Zimmerli [4], who obtained the first dislocation-free GaAs single crystals (< 15 mm diameter).

However, directional crystallization in a horizontal boat was the only practical technique for producing GaAs single crystals with an acceptable (though D-shaped) cross-section during the 1960s.

A real breakthrough in GaAs growth was made by Mullin et al. [5] in 1965 by applying a liquid boron oxide encapsulation to cover the melt; this technique was first used by Metz et al. [6] for Czochralski growth of PbTe and PbSe crystals. This liquid-encapsulated Czochralski (LEC) technology allowed a transition to reproducible production of single crystals with a circular cross-section, which is better suited for wafer production. Owing to higher thermal stress, the dislocation density of LEC-grown GaAs was higher than that of crystals grown by the horizontal-boat method. Therefore, efforts were concentrated on developing low-thermal-gradient LEC techniques such as the vapour-pressure-controlled Czochralski (VCz) [7] and fully-encapsulated Czochralski (FEC) methods [8] to reduce the dislocation density. On the other hand, the vertical Bridgman (VB) technique, first applied by Beljaka and Grishina in 1967 for III–V materials [9], was rediscovered for the growth of low-dislocation GaAs single crystals by Gault et al. [10] in 1986 and, in the meantime, has reached maturity for commercial crystal growth.

In general, single-crystal growth of GaAs (and other III–V compoundss) is more challenging than that of silicon owing some fundamentally different properties, as will be shown in the next section.

In Germany, the development from laboratory-scale manufacturing to commercialization of III–V compounds was carried out in parallel for nearly 25 years by the former Wacker Chemitronic GmbH in Burghausen and the former VEB Spurenmetalle Freiberg (together with VEB Halbleiterwerk Stahnsdorf over a period of about 15 years), both of them working in close cooperation with several universities and scientific institutes, which was essential for the success of the development. With the unification of Germany in 1990, the III–V-related activities were concentrated in Freiberg and successfully continued by Freiberger Compound Materials GmbH, which is now one of the leading GaAs wafer producers for all kinds of applications worldwide.

21.3 Properties of III–V Compounds

With the exception of the group III nitrides, which exhibit a hexagonal wurtzite structure in thermodynamic equilibrium, the remaining III–V compounds such as GaAs, InP and GaSb crystallize in the cubic zinc blende structure. In both structure types, each atom is tetragonally coordinated. The lattice parameters of some representatives of the III–V compounds are given in Table 21.1. Depending on the difference in electronegativity of the constituents, the compounds show various ionicities (heteropolarity), resulting in different band gaps. The [111] direction of the zinc blende structure and the [0001] axis of the wurtzite structure are polar axes, i.e. these directions and the opposite directions are not equivalent. A further property of these structures is piezoelectric polarization, resulting in electric fields if the crystal is strained. This has a great impact in device design for transistors and diodes.

GaAs and related compounds show a lower thermal conductivity compared with silicon, which makes the dissipation and extraction of heat from operating devices less effective, and limits the growth rate in crystallization from the melt. Table 21.1 contains data for the linear thermal expansion coefficients and thermal conductivity of the compounds.

The electronic properties of some binary III–V compounds are collected in Table 21.2. Unlike silicon, the important III–V compounds are direct semiconductors, with the exception of GaP, i.e. the conduction band minimum and the valence band maximum are at the same k -vector. This is essential for an effective conversion of electrical to optical energy and vice versa. GaAs and InP and, especially, the group III nitrides have a significantly larger band gap than that of silicon, allowing high-temperature and radiation-tolerant electronics, optical and mechanical sensors, and high-voltage rectifiers, etc. Owing to a larger curvature at the minimum of the conduction band, GaAs and InP have a lower effective electron mass, resulting in a higher electron mobility compared with silicon. Therefore, devices working at higher frequencies are possible.

Doping of III–V compounds to obtain p- and n-type conductivity in the range of 10^{-3} to $10^{-1} \Omega \text{cm}$ at room temperature, necessary for optoelectronic and microelectronic devices, is rather easily achieved. Group II elements act as acceptors, group VI elements act as donors and group IV elements show an amphoteric behaviour. Metallic and Schottky contacts can easily be produced as well. Furthermore, semi-insulating (SI) behaviour characterized by an electrical resistivity above $10^6 \Omega \text{cm}$, which is much higher than even that of intrinsic silicon, can be obtained by pinning the Fermi level at donors or acceptors in the mid-gap position. The native mid-gap double donors EL2 in GaAs and Fe at appropriate concentrations in InP and GaN cause pinning of the Fermi level under certain additional conditions. This allows the manufacturing of devices directly in the substrate without special measures to isolate the devices from each other, reducing the number of technological steps in device manufacturing.

Table 21.1. Lattice parameters and expansion coefficients of III–V compounds [11–27]

Property	Si	GaAs	InP	GaP	AlN	GaN	InN
Crystal structure	Diamond	Zinc blonde	Zinc blonde	Wurtzite	Wurtzite	Wurtzite	Wurtzite
Space group	$F\bar{d}3m$	$F4-3m$	$F4-3m$	$P6_3mc$	$P6_3mc$	$P6_3mc$	$P6_3mc$
Lattice constant a (Å)	5.431	5.65325	5.86930	5.45117	3.1106	3.1892	3.548
Lattice constant c (Å)					4.9795	5.1850	5.760
a expansion coefficient ($10^{-6}/\text{K}$)	2.59	5.93	4.75	4.65	2.9	3.5	3.6
Thermal conductivity at 300 K (W/cm K)	1.41	0.46	0.7	0.77	2.85	1.3	0.45

Table 21.2. Electronic properties of III–V compounds in comparison with silicon [11–27]

Property	Si	GaAs	InP	GaP	AlN	GaN	InN
Band structure	Indirect	Direct	Indirect	Indirect	Direct	Direct	Direct
Band gap E_G (eV) at 300 K	1.12	1.42	1.35	2.35	6.1	3.452	1.89
Effective electron mass (m_o)		0.072	0.073	0.254	0.3	0.22	0.12
Effective hole mass (m_o)		0.66	0.2	0.67	1.14	0.8	1.6
Electron mobility at 300 K (cm ² /V s)	1450	8500	5370	160		950–440	250
Breakdown voltage (kV/cm)	300	600				5000	
Static dielectric function	11.8	12.5	10.7	11.1	9.1	9.5	9.3

Table 21.3. Thermal and mechanical properties of III–V compounds in comparison with silicon [11–27]

Property	Si	GaAs	InP	GaP	AlN	GaN	InN
Melting point T_m (°C)	1410	1239	1062	1467	≈ 3000	≈ 1700	≈ 1100
Dissociation pressure at T_m (bar)	10^{-7}	2.2 (As)	27.5 (P)	32 (P)	30	2000	
Latent heat of fusion L (kJ/mol)	12.1	96.7	62.7				
Critical shear stress (MPa)	3.61	0.62	0.92	4.08			
Stacking fault energy (mJ/m ²)		55	18	41			
Fracture toughness (MPa m ^{1/2})	0.82–0.95	0.458	0.441				
Density (g cm ⁻³)	2.33	5.307	4.787	4.138	3.255	6.10	6.81

A further fundamentally important property of the III–V compounds is their mutual solubility, which allows ternary and quaternary (and higher) compounds with electronic and other physical properties roughly obeying Vegard's rule. This makes heterojunctions with tailored band structures for bipolar optoelectronic and microelectronic devices possible. Ordering is observed in ternary and quaternary compounds.

The properties related to the preparation process are exemplified by the data given in Tables 21.1 to 21.3. The melting temperatures of group III nitrides are not well defined. The III–V compounds generally decompose into their elements near the melting temperature, and therefore the vapour pressure at the melting temperature is determined by the dissociation pressure of the group V element. The vapour pressures of the group III elements are several orders of magnitude lower. So, pressure vessels are required to grow phosphides and arsenides from their melts.

The critical shear stress near the melting temperature characterizes the tendency towards thermoplastic relaxation of thermally induced stresses by the generation and multiplication of dislocations during growth and subsequent heat treatments. It is significantly lower for the phosphides and the arsenides than for silicon, which has made the growth of large-diameter single crystals of III–V compounds free of dislocations impossible until now, in contrast to silicon, where dislocation-free growth is common practice. The critical shear stress of III–V compounds is increased by doping elements, such as silicon in GaAs (solution hardening). If this is compatible with the required electrical characteristics, this property can be used to grow low-dislocation crystals. Furthermore, as is obvious from the stacking fault energies given in Table 21.3, the tendency to twinning is especially high for InP, in agreement with experience.

The resistance of a brittle material to crack propagation is characterized by its fracture toughness. The fracture toughness is lower for GaAs and InP than for silicon, for which reason breakage is a critical issue influencing yield in device manufacturing. No data are available for the group III nitrides.

21.4 III–V-Based Devices, Device Technologies, and Requirements for Substrates

III–V-based devices produced on or in III–V compound substrates can be divided into two categories [28]: optoelectronic devices and microelectronic devices. Light emitters such as LEDs (light-emitting diodes) and laser diodes, operating from the visible to the near IR, and detectors such as PIN diodes, are optoelectronic devices. They are applied in exterior/interior lighting, as emitters for optical data transfer systems, in the writing and reading of compact discs, in detectors and, to a smaller extent, in satellite solar cells. Optoelectronic devices are usually fabricated using semiconducting (to match semiinsulating) (SC) p- or n-type substrates with an electrical resis-

tivity of 10^{-3} to $10^{-1} \Omega \text{cm}$ at room temperature. All optoelectronic devices are bipolar devices. SC GaN on silicon carbide or sapphire is commonly used as a substrate for the spectral range around blue, and GaAs and InP are commonly used for the range from green to the near IR. The active layers are designed using ternary or quaternary compounds such as GaAsP, AlGaAs, InGaAlP, InGaAsP and GaAsN. In contrast, the application of silicon is restricted to pixel detectors and solar cells.

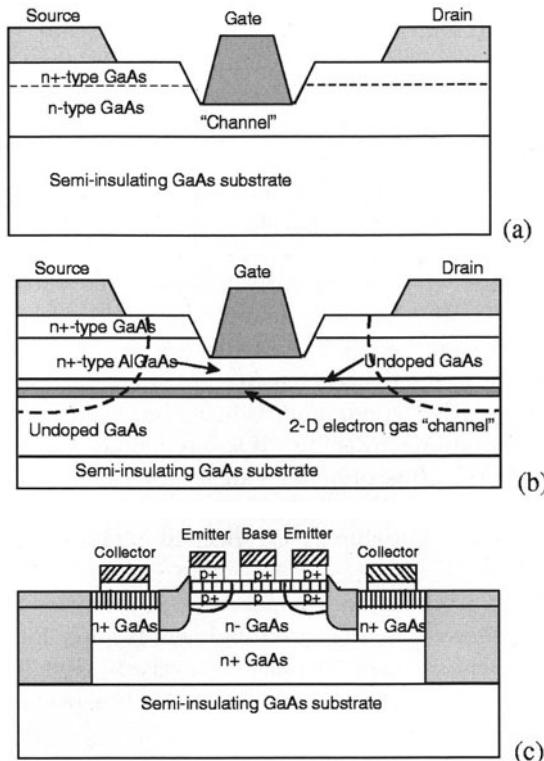


Fig. 21.1. MESFET (a), HEMT (b) and HBT (c)

Microelectronic devices are made using semi-insulating substrates, the majority of which are SI GaAs at the present time. The devices include power amplifiers and switches for mobile communication (handsets), drivers for lasers and modulators for fibre-optical communication, and low-phase-noise oscillators for millimetre-wave applications in wireless broadband communication and automotive applications (anti-collision radars, GPS, road tolling, etc.). These devices typically use MESFET (metal-semiconductor field-effect-transistor), HEMT (high-electron-mobility transistor) or HBT (hetero-bipolar transistor) structures. MESFETs and HEMTs are unipolar

devices: there are no minority carriers, and recombination is not an issue. HBTs contain p–n junctions and recombination is important as a degradation mechanism (see below). Active layers for HBTs consist of AlGaAs/GaAs, AlGaAs/InGaAs, InGaP/GaAs, etc. The basic device structures are outlined in Fig. 21.1. Owing to the better performance of HEMTs and HBTs compared with MESFETs for microwave applications, MESFET technology is predicted to lose market share to both HEMT and HBT devices.

The III–V active layer(s) of the devices are manufactured either in or on the substrate. At present roughly half of the total semi-insulating GaAs is processed using ion implantation for MESFETs; the remainder is processed by various epitaxial techniques for HEMTs and HBTs, with MBE and MOCVD holding the lead. Optoelectronic devices are exclusively produced by epitaxial methods, mainly MOCVD.

The performance of the active layer or the multilayer system is affected to varying degrees by the quality of the substrate. For example, since MESFETs are ion-implanted devices built in the substrate, the electrical resistivity and its homogeneity over the wafer and from wafer to wafer are essential for ensuring the required channel carrier concentration, which in its turn determines the turn-on threshold voltage V_{th} , the central parameter of the device. On the other hand, HEMTs and HBTs, as epitaxial devices, require epi-ready surfaces that are free of particles and have a well-defined and consistent surface oxide layer to guarantee that they can be applied without additional treatments such as etching procedures and to guarantee a high quality of the deposited layers. In addition, for HEMTs, the resistivity of the substrate can influence the turn-on voltage, and again the resistivity and homogeneity must be controlled (on both the microscopic and the mesoscopic scales) in order to avoid variations in the performance of the devices being produced. With HBTs, the most important substrate requirement is that the “leakage” through the substrate between different mesa-isolated devices should be extremely low. To achieve this, substrates with very high resistivity are used. Finally, for any bipolar device, either optoelectronic or microelectronic, the substrate should not contribute anything to the epitaxial structure that could lead to device degradation through unwanted carrier recombination.

Since both diodes (LEDs and laser diodes) and HBTs rely on injection of electrons from an n-type region into a p-type region, the degradation mechanisms are similar. For optoelectronic devices, all carrier recombination is designed to be radiative, i.e. to produce light. Any non-radiative recombination is accompanied by energy transfer to the lattice, which can cause structural defects to be generated or grow, resulting in new defects which also act as non-radiative recombination centres. Degradation continues avalanche-like until the device fails completely. Dislocations in the substrate which are continued during epitaxial growth into the layer structure (threading dislocations) are known to act as non-radiative recombination centres that reduce the lifetime of the device. Despite the application of buffer layers or special growth

methods such as ELOG, epitaxial lateral overgrowth, threading dislocations cannot be completely avoided. Therefore, for GaAs- and InP-based optical structures, low-dislocation substrates are needed. As degradation increases with the injected carrier density, the requirements concerning low dislocation density of the substrate are higher for laser diodes than for LEDs.

Low-dislocation-density GaN is not yet available, which has led to a delay in the practical application of GaN-based violet lasers with sufficient lifetime. But surprisingly, despite the high density of defects (10^{10} cm^{-2}), GaN-LEDs are nearly as effective as defect-free AlGaInP diodes produced on low-dislocation-density GaAs substrates and are widely used. This is not understood sufficiently well.

For HBTs, degradation is somewhat different. Again, non-radiative recombination at the emitter–base junction can create defects and the device ages, but this results in a gradual decay of the transistor gain, not in a sudden failure of the device as for laser diodes. This process increases with the current injection, but seems to be unrelated to any dislocations in the substrate. The mechanism is not yet well understood. Both low- and high-dislocation-density substrates have been used for HBTs until now, with a preference for low-dislocation-density substrates if high injection currents are to be applied.

For unipolar devices such as MESFETs and HEMTs, recombination at dislocations and other defects does not play a role. Therefore, the substrate type is determined by the criteria already mentioned above.

In addition to device-related requirements, cost reduction for all substrates is an everlasting task that must be carried out to preserve the competitiveness of compound semiconductors. Possibilities for reduction of the cost of substrates have recently been discussed by Bindemann [29].

21.5 GaAs: from Synthesis of the Compound to Wafers

21.5.1 Basic Considerations

A highly simplified flow diagram of GaAs wafer production is represented in Fig. 21.2. Except for the synthesis, there is no difference from the technological flow diagram for the production of Si wafers. Some underlying basic aspects will be considered in the following sections before the individual technological steps are dealt with in greater detail.

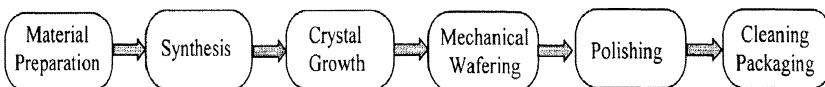


Fig. 21.2. Flow diagram of GaAs wafer production

Phase Diagram

As the synthesis, crystal growth and heat treatment of GaAs occur near to thermodynamic equilibrium, the principles underlying these process steps can be explained by consideration of the Ga–As equilibrium phase diagram depicted in Fig. 21.3.

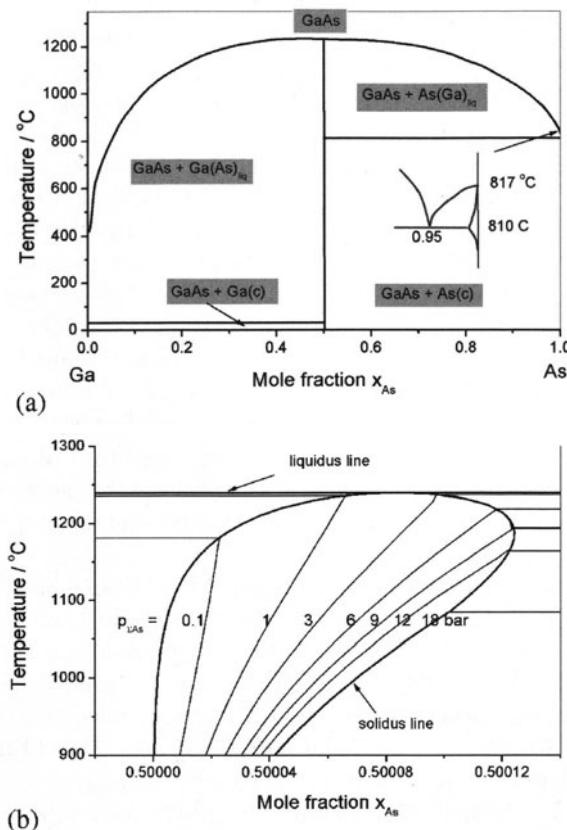


Fig. 21.3. Ga–As equilibrium phase diagram (b: detail of a)

GaAs is a congruently melting compound, dividing the phase diagram into two eutectic partial systems with eutectics close to pure arsenic and gallium and eutectic temperatures at about 810°C and 30°C, respectively [30]. Solid GaAs possesses a narrow homogeneity range. The extension and temperature dependence of the solubility (solidus) lines are not yet well established. The homogeneity range according to the best present-day knowledge is represented in Fig. 21.3b. It is asymmetric with regard to the stoichiometric composition, with a larger extension towards As-rich concentrations. The

solubility of arsenic in solid GaAs decreases with decreasing temperature. Various atomistic models including various concentrations of the possible native defects in a binary compound (namely V_{Ga} , V_{As} , Ga_i , As_i , As_{Ga} and Ga_{As}) have been developed to explain the homogeneity range. An in-depth review has been given by Wenzl et al. [30].

The isobars given in Fig. 21.3b indicate the As decomposition pressure (the sum of the partial pressures of monomeric, dimeric and tetrameric As composing the vapour phase) along the liquidus line and inside the homogeneity range of solid GaAs. The Ga partial pressure is several orders of magnitude lower and has been neglected. For congruently melting GaAs, a vapour pressure around 2 bar is commonly expected. The vapour pressure increases strongly with increasing excess of arsenic in the melt.

To avoid decomposition of the melt during synthesis and growth, two principles are in use: the “hot-wall” and the “encapsulation” principles, both already mentioned in Sect. 21.2. In the first case, the decomposition pressure of the three-phase equilibrium at the liquidus temperature of GaAs (at a given non-stoichiometry) is made equal to the vapour pressure of solid arsenic held at an appropriate temperature. At the same time, the temperature of the enclosure is kept higher than the As source temperature to avoid condensation on the walls, which gave the method its name. Obviously, by varying the arsenic pressure in the enclosure (the degree of freedom left in the system), the composition of the melt and hence that of the solid can be fixed. So, this method allows control of the non-stoichiometry by controlling the temperature of the As source in a range determined by the maximum possible As pressure in the ampoule.

In the second case, if the melt is covered with an “ideal” encapsulant that does not react with the melt, decomposition will be avoided as long as nucleation of As bubbles in the melt is suppressed by a surrounding (inert) gas pressure higher than the decomposition pressure of the melt at the composition under consideration. Therefore, at a fixed inert-gas pressure, the maximum arsenic content of the melt is defined. A local increase of the arsenic concentration above this limit, e.g. by segregation at the advancing solid/liquid interface, would result in bubble formation and a possible disturbance of the solidification. At the surface of the emerging (from the encapsulant) GaAs crystal, the surface composition is determined by the temperature and decomposition pressure. If the arsenic pressure is negligible, as in LEC, selective As evaporation takes place accompanied by a continuous change of surface composition, possibly up to the formation of Ga droplets (“Ga tears”), which must be avoided, for example by choosing an appropriate temperature distribution along the growing crystal.

It should be mentioned that the liquid boron oxide used as an encapsulant for GaAs is not an inert cover but has to be regarded as part of a complicated reaction system: Ga and As from the melt can be oxidized, incorporated into the encapsulant and partly evaporated into the surrounding atmosphere, thus changing the composition of the Ga–As melt. These reactions are controlled

by the oxygen potential in the system [31]. Correspondingly, dopants and impurities are influenced by the encapsulant as well. As an example, carbon doping will be considered in Sect. 21.5.1.

From the phase diagram, it follows that GaAs crystal growth is possible from either a Ga- or an As-rich melt. Growth from an As-rich melt is preferred for a number of reasons, such as generation of the mid-gap donor EL2 [32] as a prerequisite for semi-insulating properties, avoidance of twin formation during crystal growth [33] and better homogeneity of electrical properties [34].

If the composition of the melt deviates from the congruently melting composition, macrosegregation will take place during solidification. But as melt compositions near to the congruently melting composition are used, constitutional supercooling due to segregation of the main components is not an issue in GaAs crystal growth and the change of non-stoichiometry in the solid will be insignificant except for the very last layer to solidify [30]. Under the same conditions, microsegregation due to fluctuating solidification parameters, caused for example by natural or forced convection, would result in non-stoichiometry striations, which would be partly smeared out by diffusion; these do not lead to detectable fluctuations of native defects. Of course, when doped melts are being solidified, both macrosegregation and microsegregation must be considered.

The non-stoichiometry of the GaAs single crystal, as established at the liquid/solid interface by the melt composition, is constant inside the homogeneity range. The concentration and type of native defects in this range are determined by the deviation from stoichiometry and the temperature, as can be concluded from a thermodynamic approach (see e.g. [30]). As soon as the (retrograde) solubility line in Fig. 21.3b is reached during cooling, nucleation and growth of As precipitates occur (Fig. 21.4). Their size distribution and density follow the relationships of non-stationary nucleation theory and Ostwald ripening, i.e. they are linked to the number of nucleation sites available (homogeneous and heterogeneous nucleation), to the supersaturation (supercooling) produced, and to the resting time and the temperature below the solubility limit, which determines the average diffusion length. As a result of the generation of As precipitates, the non-stoichiometry of the solid solution (GaAs) approaches the stoichiometric composition, which results in a change of the concentration and type of “structural” native defects.

The interplay of As precipitation and the non-stoichiometry of the remaining solid GaAs is the basis for the control of the concentration and distribution of native defects by the annealing of boules under mass-conserving conditions (defect engineering [35–38]).

Semi-Insulating and Semiconducting GaAs

The semi-insulating behaviour of GaAs is basically related to the intrinsic mid-gap double donor EL2, which is most probably a single arsenic anti-

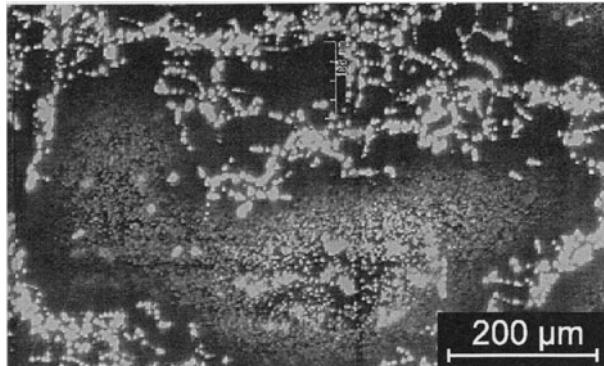


Fig. 21.4. As precipitates on dislocations and in the dislocation-free cell interior

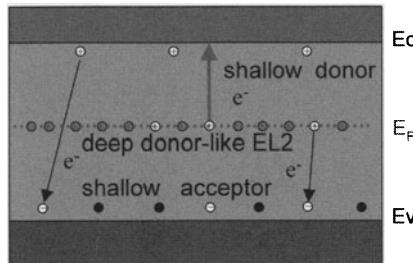


Fig. 21.5. Compensation in the case of semi-insulating GaAs

site atom (As_{Ga}) [39]. The Fermi level is pinned at the localized state of EL2 if the following condition is fulfilled: $N_{\text{EL2}^0} > N_{\Sigma A} > N_{\Sigma SD}$, where N_{EL2^0} , $N_{\Sigma A}$, and $N_{\Sigma SD}$ are the concentrations of the occupied EL2, of the acceptors and of the donors shallower than EL2, respectively. SI GaAs is n-conducting. The charge carrier concentration in the conduction band is given by $n = 4.7 \times 10^{17} [N_{\text{EL2}^0} / (N_{\Sigma A} - N_{\Sigma SD})] \exp(-E_{\text{EL2}} / kT)$, where $E_{\text{EL2}}(T) = 0.75 \text{ eV} + 2.4 \times 10^{-4} T$. In state-of-the-art SI GaAs, residual impurities are reduced to as low a level as possible and carbon is intentionally doped as a shallow acceptor. The above condition then reads $N_{\text{EL2}^0} > |\text{C}| + (N_{\Sigma A} - N_{\Sigma SD}) > 0$, with $(N_{\Sigma A} - N_{\Sigma SD}) < 10^{14} \text{ cm}^{-3}$. By keeping residual impurities low and constant, the carrier concentration and hence electrical resistivity of SI GaAs are determined by two parameters, the EL2 and carbon concentrations (see Fig. 21.5). The resistivity range from 10^6 to $10^9 \Omega \text{ cm}$ is reproducibly accessible in commercial crystal growth. Compared with silicon, the content of residual impurities in SI GaAs is several orders of magnitude higher. Pinning of the Fermi level to a deep acceptor such as Cr to obtain SI GaAs is no longer used.

The EL2 concentration increases with arsenic concentration above the stoichiometric [40]. Therefore, it can, basically, be controlled by controlling

the composition of the melt from which the solid is grown, as well as by post-growth annealing [36] using the above-mentioned solid-state reactions that occur when the retrograde solidus line of GaAs is crossed. But the variations of EL2 concentration accessible by these processes are within a factor of 2 at a maximum, and [EL2] is typically around $1.4 \times 10^{16} \text{ cm}^{-3}$.

p- and n-type semiconducting GaAs are obtained by doping with shallow acceptors (mainly Zn, Be and C) and shallow donors (mainly Si and Te), with concentrations around 10^{18} cm^{-3} , directly in the melt. The Fermi level is pinned near the conduction and the valence band, respectively. Unlike carbon, the group IV element silicon shows an amphoteric behaviour in GaAs, i.e. it can be substituted for an As as well as a Ga atom, with $[\text{Si}_{\text{Ga}}] \approx 0.8[\text{Si}]$. Furthermore, interactions between native (non-stoichiometry) defects and impurities are observed, resulting in electrically active complexes which influence compensation in SC GaAs and, therefore, have to be taken into account. Naturally, the underlying thermodynamic defect equilibria are sensitive to post-growth heat treatments.

Finally, native point defects and dopants will interact with structural defects (“gettering”) such as single dislocations and subgrain boundaries, resulting in equilibrium segregation and related local changes in, for example, the electrical properties, which are scaled by the characteristic dimensions of the dislocation structure. To give an example, according to the Petroff–Kimerling mechanism [41], interstitial arsenic As_i , which is a major native point defect in non-stoichiometric GaAs, can be absorbed by a dislocation, resulting in a single climb step, producing a Ga vacancy, which is then occupied by another As_i , generating an EL2 at the dislocation: $\text{As}_i + \text{Ga}_{\text{Ga}} = \text{GaAs}_{\text{unit climb}} + \text{V}_{\text{Ga}}$. Therefore, an enhanced EL2 concentration can be expected around dislocations, in agreement with experimental observations.

In addition to semi-insulating and semi-conducting GaAs, medium-resistivity GaAs is of some importance for certain device-manufacturing processes. It is produced by oxygen doping with $[\text{EL2}] < [\text{O}] > N_{\Sigma\text{SD}}$, resulting in a pinning of the Fermi level at this deep donor. It is characterized by $10^3 \Omega \text{ cm} < \rho < 10^4 \Omega \text{ cm}$.

Carbon Control

To accommodate a customer’s resistivity specification for semi-insulating substrates, the concentration of carbon at an essentially unchanged EL2 concentration must be actively controlled during crystal growth. This is done by controlling the chemical potentials of oxygen and carbon in the growth system, which is, from a thermochemical point of view, a complex reaction system comprising the working atmosphere, the boron oxide melt, the GaAs melt, the crucible and the graphite of the heater system. The control procedure is based on a rationalization of the redox equilibria in the reaction system by calculating *predominance area diagrams*, as suggested by Oates and Wenzl [42], instead of considering a (complete) set of reaction equations

describing the system. A predominance area diagram represents the stability region of the GaAs and boron oxide melts in a plot of $\log a_C$ versus $\log p_{O_2}$, which are taken as the two key thermodynamic parameters. The ChemSage code [43] has been used to minimize the total Gibbs free energy of the system. In brief, the GaAs melt and liquid boron oxide were described by (asymmetric) regular solutions including Ga, As, B, O, C, H, N and Si; the crucible was treated as a stoichiometric compound and the gas phase as an ideal mixture. Concentrations in solid GaAs were roughly estimated using distribution coefficients of the components. Further details of the procedure can be found in [31, 44, 45].

Since transport of reaction products in the condensed phases and reaction kinetics at the phase boundaries were not included in the equilibrium model, transient phenomena during crystal growth such as the response to a sudden increase of CO partial pressure are not described. For that purpose, an advanced segregation model has been developed by Eichler et al. [46]

Thermal Stress and Dislocations

The density and distribution of dislocations in melt-grown crystals are determined by thermoplastic relaxation of thermally and, to a much lower extent, constitutionally-induced stress by processes of high-temperature climb (associated with emission or absorption of native point defects) and glide at temperatures above $\approx 400^\circ\text{C}$. This is an experimentally and theoretically well-established result.

A comprehensive theory to calculate the density and distribution of dislocations on the basis of a given time-dependent temperature field representing the entire thermal history of the growth is not yet available (see [47] for an in-depth review). Instead, the calculation of thermoelastic stress fields for various growth states is mostly followed by qualitative arguments assuming that the higher the difference between the thermoelastic stress and the critical resolved shear stress (CRSS), the higher the corresponding dislocation density in a given volume element of the crystal. Nevertheless, this concept has been successfully applied to optimize the hot zone of the puller by a generation of crystal growers.

Since measurement of the temperature field in the hot zone of a puller is a complicated, time-consuming task and insufficient as a basis for optimization of the furnace design, numerical models for heat, mass and impulse transfer have been continuously developed from the beginning of scientific crystal growth. 2D and even 3D numerical models have now reached a level which makes them a powerful tool for tailoring the temperature fields inside the furnace and the growing crystal as a function of the configuration and dimensions of the heater and insulation.

Non-stationary convection in the fluid phases implies temperature fluctuations, resulting in variations of the growth rate and stress level around their mean levels. According to the theory of twinning during crystal growth of

III–V compounds developed by Hurle [33], temperature fluctuations at the solid/liquid/gas triple junction promote the formation of twins by increasing some critical values which are related to the stacking fault energy, which in turn is related to the energy of a twin boundary. Twinning is usually coincident with the As (V) facet. The lower the stacking fault energy, the better the temperature fluctuations must be damped out, which agrees with the experimental result that it is harder to avoid twinning in InP than in GaAs when single crystals are being grown. Again, computer modelling can help to find the appropriate conditions.

21.5.2 GaAs Synthesis

The initial process steps already demonstrate some fundamental differences between compound semiconductors and the elemental semiconductor Si. Commercially available hyperpure gallium and arsenic, supplied in bottles filled with pure nitrogen to avoid oxide formation, which usually do not need further purification are used as the starting material. Two methods are practised to synthesize the compound from its constituents: low- and high-pressure methods.

The low-pressure method has already been partly described in Sect. 21.2. The elements are placed in quartz boats, which are sealed in a quartz ampoule and positioned in a two-zone furnace in such a way that the arsenic is held at a constant temperature around 620°C, necessary for a sublimation pressure > 1 bar, and the gallium is in a shallow temperature gradient with a starting temperature slightly above the melting temperature of the compound at one end of the boat (Fig. 21.6). The temperature gradient is then moved along the Ga boat and the charge is directionally solidified, avoiding bubble formation by proper temperature control. The composition of the ingot is determined by the arsenic vapour pressure and can be controlled. Instead of a stationary quartz boat in a moving temperature field, graphite boats with local rf heating have also been used.

Owing to the reaction of Ga in the melt with the SiO_2 of the quartz wall, the volatile suboxides Ga_2O and SiO are formed and Si penetrates into the melt. If these suboxides diffuse to colder parts, SiO will condense and Ga_2O will react with gaseous arsenic to form solid Ga_2O_3 ($3\text{Ga}_2\text{O} + \text{As}_4 = \text{Ga}_2\text{O}_3 + 4\text{GaAs}$). Therefore, the reaction at the melt/wall contact will continue, which increases the contamination of the melt with silicon and causes wetting and adherence of the charge to the quartz boat. To suppress these processes, diffusion and condensation of the suboxides must be prevented. This is done by a diffusion barrier between the two zones and by keeping the temperature in the hot zone high enough [27].

Owing to unavoidable Si contamination when quartz boats are used, the method is used mainly to synthesize Si-doped GaAs. Before the era of carbon-controlled SI GaAs, Cr–O-doped SI GaAs and O-doped medium-resistivity GaAs were synthesized and grown by HGF, horizontal gradient freeze. It

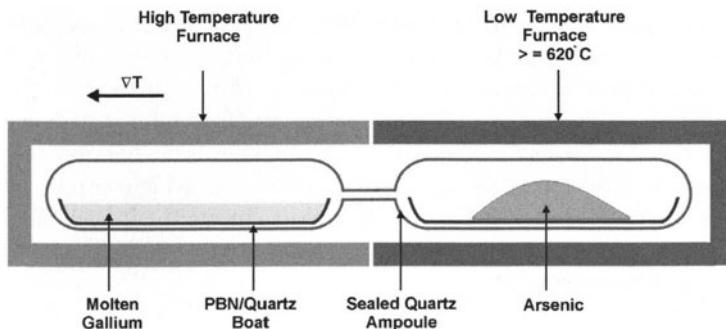


Fig. 21.6. Gallium arsenide synthesis by horizontal-gradient freezing (taken from [48])

should be mentioned that the boron content of the synthesized charge is very low. Boats made of pBN are available, too. Typical weights of synthesized bars are around 5 kg.

High-pressure synthesis from the components is performed either in separate equipment (ex situ synthesis) or combined with subsequent growth in a single-step process (called the direct-synthesis method) immediately afterwards in the puller. A pBN, pyrolytic boron nitride crucible is charged with the components in the stoichiometric composition, covered with a pellet of boron oxide and placed in a high-pressure, high-temperature vessel. Argon and nitrogen are used as working gases. During heating up, the boron oxide forms a viscous melt at around 450°C and coats the charge before excessive As sublimation would cause an As loss. The exothermic reaction of the components starts at about 800°C, accompanied by a temperature rise. The pressure reaches 70–80 bar at the melting point of GaAs. To avoid As gas bubble formation before the reaction starts, the working gas pressure is set higher than the As critical pressure at the beginning. After a homogenization period slightly above the melting temperature, either the charge is cooled to room temperature or the starting temperature is set, the chamber pressure can then be decreased and crystal growth initiated by seeding and pulling the crystal.

Composition control is more difficult for this method than for HGF. To maintain a nearly stoichiometric or slightly As-rich melt, a slight excess of As is added to compensate for uncontrolled As loss during the heating-up cycle. The excess is empirically determined and kept constant by a reproducible and reliable technology.

Carbon, oxygen and boron essentially follow the relationships mentioned in Sect. 21.5.1. But in the case of a two-step procedure, the homogenization period above the melting temperature of GaAs during synthesis is not sufficient to establish thermodynamic equilibrium in the entire system, allowing active carbon control. This is due to the slow transport of species in

boron oxide, which is mainly diffusive and decouples the gas phase from the GaAs melt. Therefore, the oxygen and carbon potentials are determined by the water content in the boron oxide and the carbon content of the melt, introduced by carbon contamination of the arsenic. Thus, a reduction of the carbon content relative to its initial value, depending on the water content of the boron oxide, is realized, resulting in a rather well-defined initial state for active carbon control during subsequent crystal growth. Furthermore, the water content of the boron oxide causes a “gettering” of impurities.

21.5.3 Crystal Growth

Crystal growth methods of commercial interest for GaAs can be classified into two groups: pulling techniques and crucible techniques. The LEC and VCz methods are GaAs-related modifications of the Czochralski method with the same general features as for silicon, and belong to the first group. The horizontal and vertical Bridgman (HB and VB) and horizontal and vertical gradient freeze methods (HGF and VGF, vertical gradient freeze method) belong to the second group. Crucible techniques are not applied for silicon. In the Bridgman method, the charge is directionally solidified by moving the crucible and/or the furnace (in opposite directions), in the HGF and VGF methods the temperature field is moved along the charge by controlling the heater power. This is schematically shown in Fig. 21.7 for vertical growth.

The crystal diameter in Czochralski growth is controlled by meniscus stability. This implies radial temperature gradients, which are basically not necessary in crucible methods. Therefore, pulling techniques are characterized by higher temperature gradients in the melt and the crystal, accompanied by higher thermal stress in the solid.

Furthermore, the axial temperature gradient and the gravity vector are antiparallel in the VB and VGF methods but parallel in the LEC and VCz methods, thus promoting natural convection in the melt in LEC growth. Together with higher temperature gradients, more vigorous (non-stationary) convection in the melt will exist, which can affect the shape of the solid/liquid interface, the instantaneous growth rate and the three-phase junction at the meniscus. As all these factors will influence structural quality, much effort has been directed to controlling melt convection in Czochralski growth, e.g. by forced convection caused by rotation of crucible and/or crystal. For the same reason, strong non-stationary convection will exist in the gas phase, resulting in surface temperature fluctuations in the growing crystal, for example. In the HB and HGF methods, the temperature gradients are mainly perpendicular to the buoyancy and convection will start without any threshold.

As already discussed in Sect. 21.5.2, wall/charge contact is an issue to be considered in all crucible methods. It not only potentially reduces the purity of the charge (an effect that occurs in pulling techniques too), but can destroy single-crystal growth by heterogeneous nucleation at wetting points, thus limiting yield. This is a reason why pBN crucibles and boron oxide

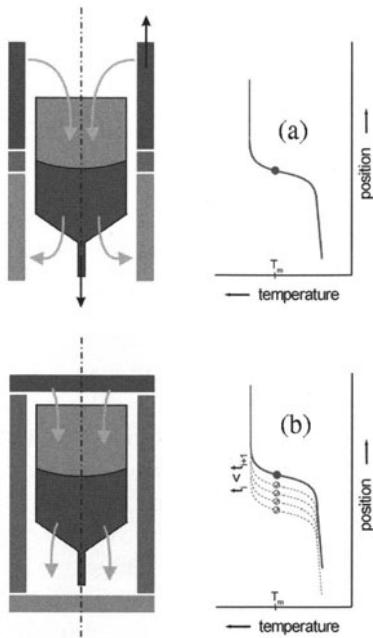


Fig. 21.7. Schematic representations of (a) VB and (b) VGF (Stöber/Tammann-type) methods

encapsulation are widely utilized with the VB and VGB methods at present; the use of pBN crucibles and boron oxide encapsulation is also a simpler technology than the possible alternatives.

Despite a rather high reduced melting entropy (Jackson factor, $\alpha \approx 6$) the $\{100\}$ solid/liquid interface is assumed to be atomically rough, whereas $\{111\}$ is flat, requiring a larger kinetic undercooling at a given growth rate [49]. Therefore, faceting is observed in regions of strong deflection of the interface, even for the $\langle 100 \rangle$ growth direction. This is especially observed at the three-phase junction and the cone-shaped part of the crystals, manifesting itself as “Ga and As facets” on the crystal surface. Owing to the polar character of $\{111\}$, the widths of the facets (the intersection of the facet with the crystal surface) are different. It should be mentioned that twinning is related to the existence of $\{111\}$ facets.

A comparison of the growth techniques mentioned above is represented in Table 21.4 [48]. Obviously, no all-purpose crystal growth technique currently exists.

Concerning the share of the different methods of the overall crystal growth capacity, for 2001 (the year of a dramatic decline in GaAs wafer production) the following numbers have been given [50]: 56% of SI GaAs wafers were produced from LEC-grown crystals, the remainder were produced by the

Table 21.4. Comparison of growth methods for GaAs [31]

	HB/HGF	VB/VGF	LEC	VCz
Temperature gradient	Low	Low	High	Middle
Solid/liquid interface control	Poor	Good	Good	Good
Diameter control	Fair, D-shaped	Good	Fair	Fair
In situ observation	Good	Poor	Good	Good
Convection	Low	Low	Strong	Fair
Maturity of technology	Good	Good	Good	Fair

VGF and VB methods, but a continuous increase in VB and VGF is expected over the next few years. 61% of the crystals were 100 mm in diameter, 34% 150 mm, and the remainder 3 and 2 inch. An increase in 150 mm diameter crystals is expected. The market is shared between five main players only.

A different situation is observed for SC GaAs: 4% of the crystals were produced by LEC, 47% by VGF and VB and 49% by HB. By diameter, the shares are 80% 2 inch, 10% 3 inch and 7% 100 mm. The 2 inch size will remain the dominant diameter in the future. The market is shared between four main players at present.

Technical and technological details of industrial pullers are rarely published. In the following, some general features will be given.

Czochralski Method

To accommodate the demands for longer and thicker crystals, and to improve the economics of crystal growth, a new generation of high-pressure pullers has been developed and is now being used in the III-V industry (Fig. 21.8). A thorough review containing the technical developments up to 1994 can be found in [51].

High-pressure vessels (up to 100 bar, suited for direct synthesis), pressure-tight feed-throughs for translation and rotation of the crystal and crucible, and a multi-heater system (usually three heaters) made of graphite are characteristic. In the case of a three-heater set-up, the central heater is used for diameter control, and the lower sub-heater defines the axial temperature gradient in the bottom region of the crucible. The upper sub-heater controls the temperature at the crystal surface to prevent decomposition by selective arsenic evaporation. The pullers are designed for crucibles up to 16 inches in diameter, allowing charges up to 50 kg [52]. pBN is exclusively used as the crucible material. The ambient gas pressure is usually between 2 and 20 bar and the gas consists of N₂ or Ar. For C, B and O control, a gas management system is available. The pullers are equipped with a fully computerized process and diameter control system.

In a difference from the practice for silicon, diameter control in GaAs Czochralski growth is based on a continuous measurement of the crystal

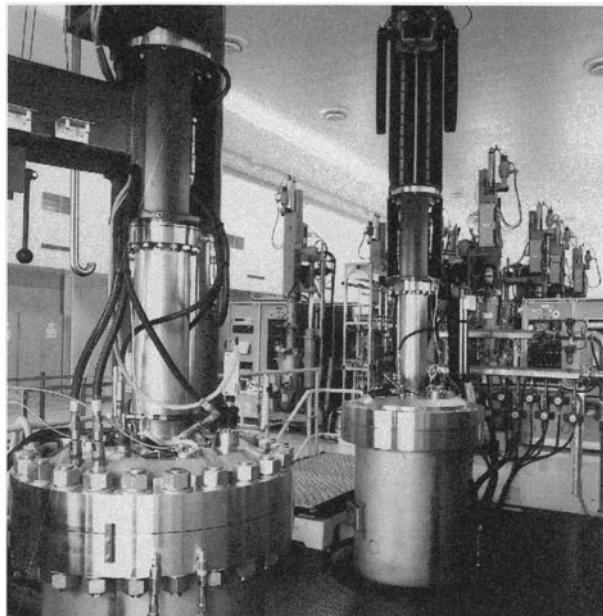


Fig. 21.8. New-generation LEC pullers

weight and its time derivatives by a weight cell integrated into the pulling head. To prevent uncontrollable temperature fluctuations in the melt driven by buoyancy forces, the melt height should not exceed a crucible-diameter-dependent height. Counter-rotation of the crystal and crucible and an optimized temperature of the lower sub-heater are used to partly damp the turbulent convection. Crystals up to 200 mm in diameter [53] and long crystals up to 350 mm with a diameter of 150 mm [54] have been successfully grown.

The growth rate of the LEC method is typically between 5 and 14 mm/h with a $\langle 100 \rangle$ seed orientation. Necking is used, but owing to the low CRSS, it is not suited to growing large-diameter single crystals. Axial temperature gradients are in the range of 80–150 K/cm. The solid/liquid interface in LEC growth is convex (as seen from the melt). According to [55], it is essential to avoid concave parts of the interface as they will lead to an agglomeration of dislocations and possibly to a polycrystalline transition.

Vapour-pressure-controlled Czochralski growth is thoroughly reviewed in [56]. The method is commercially applied for InP [57] but has not been applied for GaAs until now.

VB/VGF Method

Compared with the LEC method, practically no information about the furnace structure of commercially used pullers and the growth technology is

available. Total and partial encapsulation of the melt seems to be used, which would allow stoichiometry control at a minimum, as can be surmised from [58].

A Tammann–Stroeber- type furnace with a heater system similar to that first described by Ramsperger and Melvin [59] (see Fig. 21.7b) and scaled up from a design given in [60] is used by FCM, Freiberger Compound Materials GmbH. The temperature field and its upward movement at a given rate are preformed by an appropriate highly stable power control of the heaters. The furnace is surrounded by heat insulation and housed in a pressure vessel designed for pressures up to 10 bar. It is equipped with a computerized process control system, which includes a gas management system for control of total and partial pressure. Thus, carbon control can be attained in VB/VGF growth of GaAs [45].

Crystals are grown under total encapsulation in typical cone-shaped Bridgman crucibles made of pBN, which are heat-treated prior to usage [61]. The polycrystalline ingots, with dimensions corresponding to the crucible, are produced by high-pressure synthesis as described above. pBN has a lamellar structure and is wetted by liquid boron oxide, and this film protects GaAs from decomposition and retards interaction with the crucible. Furthermore, as the boron oxide solidifies it adheres to the GaAs, and an annular space between the GaAs and the pBN wall can be formed owing to its lamellar structure. This space protects the crystal from stress [62].

The temperature gradient in the solid at the liquid/solid interface is adjusted to be 3–5 K/cm, and the growth rate is between 2 and 4 mm/h, i.e. significantly lower than for LEC growth. Calculation of the stress level in the crystal has revealed a value of 2 MPa in the cone-shaped part of the crystal and below 1 MPa in the cylindrical part.

Crystals up to 150 mm in diameter are produced on a large scale by this technique. Only recently, SI GaAs single crystals 200 mm in diameter have been grown, showing that further scaling up is basically possible [63].

A drawback of the VB and VGF methods is the difficulty in controlling the seeding process. This has been overcome by a reproducible and very stable control of the temperature profile.

HB/HGF Method

A rather recent overview has been given by Tatsumi and Fujita [27], including technically and technologically relevant information. This method is exclusively used for Si-doped GaAs.

According to [27], a highly sophisticated furnace set-up has been developed with three zones for the three temperature levels required. Each of these zones is divided into subsections, not only longitudinally but also in the cross-section. The three zones are kept at 1240–1247°C, 1120–1220°C and 600–620°C (for the As source). After seeding, the temperature field of the hottest zone is moved to the end of the boat, keeping the grown crystal

in the second zone. Quartz is obviously used as a crucible material. This temperature distribution and a diffusion barrier between the As and crystal zones prevent the crystal from sticking to the boat, as already discussed above.

It should be pointed out that the longitudinal temperature gradient has to be controlled to between 1 and 3 K/cm and the growth rate between 3 and 6 mm/h in order to grow low-dislocation-density GaAs crystals.

Typically, $\langle 111 \rangle$ -oriented seeds are used. Therefore, $\{100\}$ wafers generally have a D-shape, unless they are ground into round wafers, which causes significant material loss. This is a disadvantage of the method, but crystals up to 1000 mm in length, allowing about 500 round 2 inch wafers, can be grown. The method is able to produce single crystals with shoulders of 85 mm and even 110 mm, suitable for 3 inch and 100 mm wafers.

21.5.4 Heat Treatment

Crystals are generally heat-treated after growth. Depending on the main target, various single-step and multi-step time-temperature procedures have been developed and applied. The processes which can be used during a heat cycle are schematically shown in Fig. 21.9: stress relaxation in the heating-up step by local dislocation annihilation and rearrangement, dissolution of As precipitates if a temperature/composition inside the homogeneity region is reached, local homogenization followed by a reprecipitation of As when the solidus line is crossed during cooling down, and establishing a new near-to-equilibrium state of native point defects at a holding temperature before final cooling to room temperature. These processes allow defect engineering.

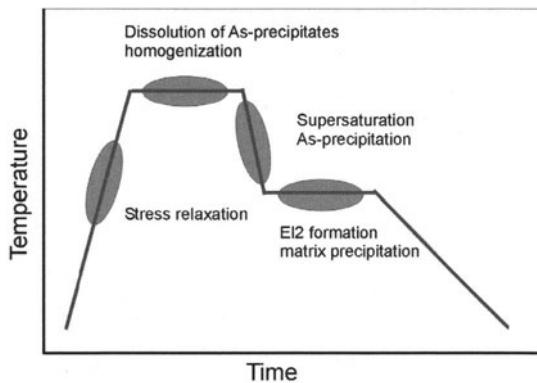


Fig. 21.9. Processes used in boule annealing

Stress relaxation is mandatory for LEC-grown crystals to avoid wafer breakage during subsequent mechanical wafering. Low-dislocation-density crystals grown by low-gradient techniques do not require a stress relaxation

procedure. In contrast, for these crystals the annealing procedure (if necessary and applied at all) must be organized in such a way that an increase of dislocation density is excluded, i.e. by a proper choice of heating and cooling rates, keeping in mind the low heat conductivity of GaAs.

Boule annealing is performed in evacuated, pretreated, sealed quartz ampoules under an As overpressure to prevent surface decomposition if a long-term soak in the homogeneity range is applied.

FCM uses a multi-step procedure including a dissolution and homogenization period at high temperatures for SI LEC crystals, with special requirements concerning mesoscopic homogeneity of electrical properties and a low stress level, but uses a medium-temperature soak for SI VB/VGF crystals to improve and stabilize the native-point-defect structure. SC VB/VGF crystals are not post-growth annealed.

Wafer annealing under a defined As partial pressure has been shown to influence the stoichiometry-related native-defect concentration in a near-surface layer by dissolving As precipitates [64]. There is no definite information on whether wafer annealing is used in practice or not.

21.5.5 Assessment of Crystals

The grown crystals are subjected to an intensive inspection programme, which is typically performed on seed- and tail-end test wafers from each crystal, and routinely includes structural evaluation after KOH etching and measurement of resistivity, carrier concentration and mobility. The etch pit density (EPD) is measured to assess dislocation density according to ASTM F1404 for LEC crystals, but for low-dislocation-density crystals, FCM applies a 100% area mapping performed by specially developed high-efficiency equipment to obtain reliable and comparable EPD data. In addition, for SI material the radial distributions of the resistivity is measured by TDCM (time-dependent charge measurement) topography [65] and EL2 concentration by near-infrared absorption. The carbon and boron contents are measured by FTIR, Fourier transform infrared spectroscopy, respectively.

Spot checks are made to trace the mesoscopic resistivity distribution by high-resolution TDCM and EL2 mapping, as well as by PCT (point contact current) measurements [66] for SI GaAs. As precipitates in SI GaAs are studied by LST (laser scattering tomography). The impurity content is followed by GDMS, glow discharge mass spectroscopy analyses, and the concentration of dopants by ICP MS, inductively coupled plasma mass spectrometry and ICP AES, inductively coupled plasma atomic emission spectrometry. Finally, the structural quality is additionally checked by double-crystal rocking-curve mapping and the residual stress level by IR polariscopy [67]. Breakage is investigated by a biaxial test described in [68].

Structural Properties

Typically, for LEC-grown material, the dislocations are arranged in a very pronounced cellular structure with globular cells in the central and peripheral regions and elongated cells in between, corresponding to a weak W-shaped etch pit density. The average EPD is about $5 \times 10^4 \text{ cm}^{-2}$ and $1 \times 10^5 \text{ cm}^{-2}$ for 100 mm and 150 mm diameter SI GaAs crystals, respectively. The EPD in the $\langle 100 \rangle$ radial direction is slightly higher than in the $\langle 110 \rangle$, which is related to the orientation-dependent Schmid factor controlling shear stress in the glide systems. The cell size is in the range 0.2–0.4 mm. With increasing EPD (at increasing crystal diameter), the cell size decreases. Te doping causes a weak reduction of EPD.

As a consequence of lower thermal stresses during growth and cooling, the average EPD of VB/VGF crystals is lower by an order of magnitude or more compared with LEC-grown crystals. The cell size is much greater ($> 1\text{--}2 \text{ mm}$) and the dislocation density in the cell walls is smaller. In a ring between the centre and the periphery, the cell walls are more or less fragmented.

Owing to solution hardening, the dislocation density of Si-doped VB/VGF-grown crystals is significantly lower than in VB/VGF-grown SI GaAs, and depends on the Si concentration. Cells composed of dislocations are no longer observed. For an average EPD below 100 cm^{-2} , most of the remaining dislocations are concentrated in the centre of the wafer and along $\langle 110 \rangle$, which is probably caused by the dominance of single glide processes up to higher temperatures.

The dislocations in SI GaAs are decorated by As precipitates, the size distribution of which is shifted to greater characteristic dimensions with decreasing dislocation density. As dislocation precipitates are present in as-grown crystals, it follows that the formation and rearrangement of dislocations are already finished when precipitation starts. In contrast, dislocations involved in slip line generation are not decorated by arsenic precipitates or are only partly decorated. Matrix precipitates, usually absent after growth, can be initiated in LEC-grown crystals by an appropriate boule-annealing procedure including a quick cooling after homogenization [69]. No precipitates are observed in Si-doped crystals.

High-resolution X-ray rocking-curve mapping ($1 \times 1 \text{ mm}^2$, footprint $0.5 \times 2 \text{ mm}^2$, Cu K_α , $\langle 110 \rangle$ rocking axis and $\{004\}$ reflection) was performed to assess local and global lattice misorientation connected with the dislocation structure. The average FWHM (full width at half maximum) for VGF SI GaAs material ($10.0 \pm 0.2 \text{ arcsec}$ and $10.5 \pm 0.3 \text{ arcsec}$ for 100 mm and 150 mm diameter crystals, respectively) is significantly lower than that for LEC wafers ($13.5 \pm 1 \text{ arcsec}$ and $16 \pm 2 \text{ arcsec}$), reflecting their lower average dislocation density and higher structural perfection.

The stress analyser SIRDTM [67], which exploits the effect of stress-induced birefringence, has been applied to estimate the global and local residual stress in the wafers. Typically, the difference between the principal

stresses in the wafer plane, used as a measure of residual stress, is between 0.1 and 0.4 MPa for 150 mm SI VGF GaAs and is slightly higher, between 0.3 and 0.7 MPa, for LEC wafers. The annealing of VGF crystals, performed to improve homogeneity of electrical properties, does not change this stress level. The residual stress level increases slightly with crystal diameter for both VGF- and LEC-grown crystals.

On the average the level of residual stress in the wafers is lower than the critical shear stress for plastic deformation at the temperatures which are used during device manufacture (for epitaxial growth and for activation annealing after ion implantation). Owing to the overlapping ranges of the residual stress levels of VB/VGF and well-annealed LEC materials, the often-assumed greater sensitivity of LEC wafers to slip line generation during heat treatment is not justified in general.

Global Electrical Properties

Improvements in raw materials, state-of-the-art cleanliness in material handling, and progress in understanding of the thermochemistry and of species transport during synthesis and crystal growth have led to a significant reduction of extrinsic defects (impurities), and to a transition from the growth of so-called “nominally undoped” to carbon-controlled semi-insulating GaAs crystals, by both the LEC and the VGF methods.

The electrical resistivity versus carbon concentration of 150 mm LEC and VB/VGF crystals is represented in Fig. 21.10. Two resistivity regions can be distinguished: semi-insulating GaAs for $[C] > \approx 2 \times 10^{14} \text{ cm}^{-3}$, where $\rho \propto [C]$ (see Sect. 21.5.1), and medium-resistivity GaAs for $[C] < \approx 1 \times 10^{14} \text{ cm}^{-3}$, with a vanishing dependence on carbon concentration. As mentioned above the Fermi level is pinned by the EL2 and O_{As} levels in semi-insulating and medium-resistivity GaAs, respectively. According to Fig. 21.10, the electrical resistivity of VB/VGF GaAs is slightly higher than that of LEC materials, especially for $[C] < 1 \times 10^{15} \text{ cm}^{-3}$. This is related to the lower average EL2 concentration of VB/VGF GaAs (see below) and, in addition, possibly to an intrinsic acceptor such as a Ga vacancy. The $\rho/[C]$ relationship holds for other crystal diameters as well. The solid line in Fig. 21.10 has been calculated using the charge neutrality condition. At $[C] > 5 \times 10^{15} \text{ cm}^{-3}$ a slight increase of the slope of the $\rho/[C]$ relationship is observed, which has been explained by an increase of EL2 concentration with increasing C content [70].

In Fig. 21.11, the Hall mobility is presented as a function of electrical resistivity for the range shown in Fig. 21.10. The decrease of mobility with increasing carbon concentration after the broad maximum around $2 \times 10^7 \Omega \text{ cm}$ is caused by scattering of electrons by the increasing concentration of ionized defects (mainly EL2^{+/++} and C_{As}⁻), whereas the deep depletion of the mobility for $\rho \approx 1 \times 10^6$ to $1 \times 10^4 \Omega \text{ cm}$ is related to mesoscopic inhomogeneities, with a gradual transition from semi-insulating to medium-resistivity behaviour starting at the cell boundaries [71].

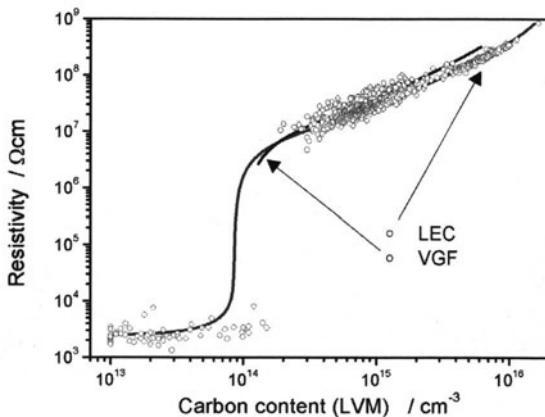


Fig. 21.10. Resistivity of 150 mm crystals as a function of carbon concentration

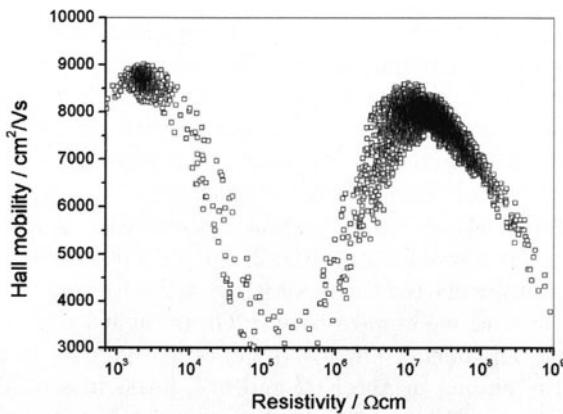


Fig. 21.11. Hall mobility as a function of electrical resistivity for 100 mm LEC GaAs

Average EL2° concentrations for boule-annealed low-carbon LEC- and VB/VGF-grown SI crystals are about $1.5 \times 10^{16} \text{ cm}^{-3}$ and $1.2 \times 10^{16} \text{ cm}^{-2}$, respectively. There is no influence of crystal diameter on the average EL2 concentration, but the spread of the EL2 concentration from crystal to crystal depends on the dislocation density and cell size. Finally, for Si-doped VGF GaAs, a linear relationship between carrier concentration and Si content has been found for $[\text{Si}] < 2 \times 10^{18} \text{ cm}^{-3}$, in agreement with current literature [72]. The deviation for $[\text{Si}] > 2 \times 10^{18} \text{ cm}^{-3}$ is due to an increasing compensation of the Si_{Ga} donor by Si_{As} and the $\text{Si}_{\text{Ga}}-\text{V}_{\text{Ga}}$ complex [73], both acting as acceptors and increasing with the Si content of the crystal. Similar relationships hold for Te-doped GaAs.

Homogeneity

When one is discussing homogeneity of GaAs crystals, it is necessary to distinguish between several different length scales. Macroscopic, mesoscopic and microscopic homogeneity can be defined as types of homogeneity that scale by crystal length/crystal diameter, cell size, and distance between individual dislocations, respectively. Macrosegregation and microsegregation at the solid/liquid interface and equilibrium segregation in the solid state are the underlying mechanisms of these inhomogeneities. Therefore, to control macroscopic homogeneity, the melt composition ahead of the interface must be controlled. The corresponding measures are reduction of the impurity concentration to a level, dependent on the distribution coefficients, such that the concentration incorporated is below a certain threshold; control of carbon concentration by appropriate carbon and oxygen potentials in the gas phase; and keeping the melt composition near to the congruent melting point. In addition a flat solid/liquid interface would help to avoid radial inhomogeneities in the C concentration. On the other hand, mesoscopic homogeneity can be influenced by solid-state diffusion during boule annealing.

For evaluation of the macroscopic homogeneity across wafers, TDCM mapping is performed with a lateral resolution of a few millimetres, therefore averaging over individual cells. An example is given in Fig. 21.12 for a 150 mm wafer taken from an LEC (left) and a VB/VGF (right) crystal. The standard deviation of the electrical resistivity is $<20\%$ for LEC and $\approx 30\%$ for VB/VGF crystals, respectively.

The standard deviation of the radial $EL2^\circ$ concentration is typically around 3.0% for 150 mm LEC crystals, compared with about 4.5% for 100 mm

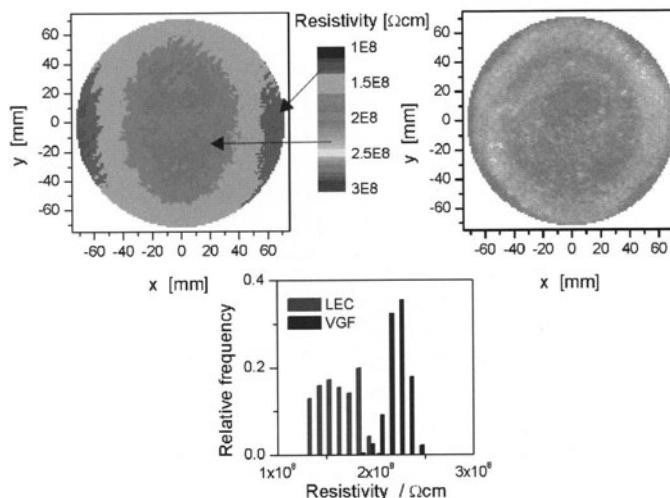


Fig. 21.12. Typical TDCM maps of 150 mm LEC (left) and VGF (right) wafers

material. The typical axial homogeneity of 150 mm SI GaAs with active carbon control is better than 20% for both growth technologies.

Microsegregation of impurities and of carbon and non-stoichiometry, due to a fluctuating growth rate or momentum boundary layer, is of minor importance in the LEC and VB/VGF growth of SI GaAs crystals.

The axial carrier concentration of Si-doped SC GaAs is governed mainly by macrosegregation of Si, which cannot be influenced, because growth takes place under mass-conserving conditions, unlike the growth of C-doped crystals.

To evaluate the mesoscopic homogeneity, EL2°, TDCM and PCT mapping is performed with a lateral resolution resolving individual cells. Typically, an enhanced electrical conductivity is found in the cell wall region, correlated with an enhanced EL2° concentration. Mesoscopic homogeneity can be improved by boule annealing, resulting in an improvement or at least change of the global properties of the crystals. Thus, the Hall mobility increases after mesoscopic homogenization for both LEC and VB/VGF GaAs.

Wafer Breakage

Wafer breakage is another big issue that is sometimes discussed in connection with global residual stress due to radial inhomogeneities in the dislocation distribution [74]. To elucidate this matter and, additionally, the possible influence of the surface quality of the wafers, fracture strength tests have been introduced. The “ring-on-ring biaxial strength test” according to DIN 52 292 [68] has been modified such that a small-diameter ball (1/8 inch steel) is used to load the wafer centrally. The load–displacement curves are recorded, and the maximum load and bending at fracture are the key parameters for evaluation. Since the maximum bending at fracture is much larger than the wafer thickness, numerical solutions of the von Karman theory have been used to calculate the fracture strength [75], taking the Hertz-type loading into account. Statistical evaluation of the data has been done by means of Weibull distribution analysis.

As an example, Weibull plots of the load at fracture for two sets of 150 mm LEC and VGF wafers are represented in Fig. 21.13. The tests were performed with the final polished and cleaned surfaces of the wafers in tension. The inserts indicate the threshold load F_o , the 63.2% probability Weibull fracture load F_c and the Weibull coefficient. Slightly higher loads at fracture are observed for LEC wafers. As LEC wafers have a higher residual stress level on average, residual stress is obviously not the main factor limiting fracture strength.

By studying the influence of different etching and polishing procedures on wafer breakage, it has been concluded that surface defects are the main reason for early wafer breakage. As expected, laser marking reduces the load at fracture dramatically, with soft laser marking being the best solution [76].

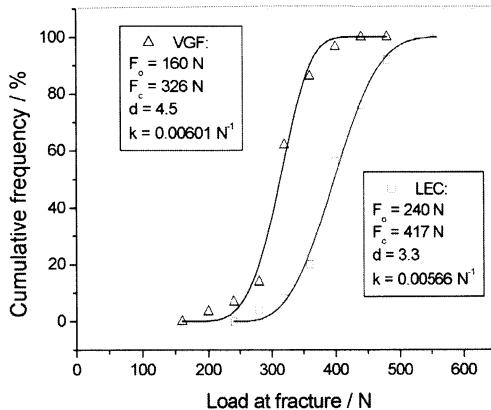


Fig. 21.13. Weibull plots of load at fracture for 150 mm LEC and VGF wafers

21.5.6 Wafering

Wafering involves a series of precise mechanical and chemical processing steps that turn a segment of the ingot into a functional wafer. It includes cropping of the seed and tail ends, taking samples for material testing, grinding of a cylindrical rod to a specified diameter, grinding a flat or a notch, oriented slicing, edge grinding, optional lapping or surface grinding, laser marking, damage etching, polishing and surface conditioning, combined with final cleaning and packaging. Each step is designed to bring the wafer into compliance with a customer's specification. There is no difference between the wafering behaviour of VB/VGF GaAs crystals and that of LEC material.

The processing steps and the equipment used correspond in general to those used for wafering of silicon, but the different physical, mechanical and chemical properties of GaAs, as well as crystal-growth-related peculiarities, have led to modifications of the equipment and differences in the technology.

As device-manufacturing technologies based on ion implantation and epitaxy require different specifications, wafering is organized not by batch but rather by crystal for both semi-insulating and semiconducting GaAs. In addition, wafering maintains the true position of the wafer in the boule, with the front and back surfaces of the wafer directed towards the seed and tail ends, respectively, of the crystal. This procedure is applied from slicing to packaging and requires a strong logistic effort, as wafer tracking must be possible at each processing step.

Two systems of flats for orientation definition are applied (US/SEMI standard and Europe/Japan specification), which do not have a relation to the conductivity type or wafer orientation. For laser production, an extremely high precision of the orientation of the primary flat of SC substrates ($\{110\}$ with an error $< \pm 0.1^\circ$) is required, as cleavage is used to manufacture the

optical confinement of edge-emitting lasers. GaAs wafers 150 mm in diameter (and larger) are notched instead of flattened, which reduces the dynamic imbalance in spinning procedures, this imbalance being more pronounced in GaAs owing to its higher density. In addition, for epitaxial overgrowth, a great variety of off-orientation (0.4–15°) specifications exist, requiring a high flexibility in the wafering process.

For wafer diameters of 100 mm and greater wafers are nearly always polished on both sides in order to reduce wafer breakage during device manufacturing. And finally, wafer surfaces must be conditioned to be “epi-ready” so that they can be loaded into the equipment for epitaxial overgrowth without preliminary wet-chemical treatments. This again requires a tight control of the corresponding processing steps.

Mechanical Wafering

Slicing of GaAs ingots is carried out by inner-diameter (ID) sawing and, more recently, wire sawing for diameters ≥ 150 mm. Owing to the higher susceptibility of GaAs to wafer breakage, the feed rates for ID sawing are lower (8–25 mm/min) than those for silicon, which are typically 60 mm/min. Warping and waviness of cut wafers mainly are caused by deflection of the blade or wire during the cutting process. Therefore, lapping and surface-grinding steps are often used to improve the quality, but these steps create additional material losses and add costs. This problem has initiated detailed studies of the forces acting on the blade/wire, regarding not only “technological” forces caused by external process conditions as in the past, but intrinsic forces too, which are mainly determined by the material properties of the semiconductor being cut [77]. It has been recognized that crack formation, the elementary process for the cutting or grinding of brittle materials, is assisted by dislocations and therefore is orientation-dependent, owing to the existence of two types of dislocations in as-grown and annealed GaAs, namely α - and β -60° dislocations. A new approach to high-quality wire sawing and polishing has been developed, which, in addition to an increase in feed rates of up to 400% compared with standard processes, has resulted in improved warping and waviness of the cut wafers, making lapping and surface grinding no longer necessary.

The damage depth of the cut wafer surface is significantly smaller for wire sawing than for ID sawing, requiring less material removal by subsequent damage-etching procedures.

The edges of as-cut wafers are sharp and require shaping in a customized way, which is done by an edge-grinding operation where grinding wheels are used, the profile of which is determined by the target wafer. Depending on the final edge contour or roughness specifications, varying sizes of grinding grains are used. There are problems inherent in this process which have led to alternative edge-shaping procedures.

Polishing and Final Cleaning

Compared with silicon, more reactive agents are necessary for mechano-chemical polishing of GaAs. As a consequence, the thickness to be removed will be greater, facilitating the generation of thickness variations between the periphery and the centre of the wafer due to the complicated kinematics of the wafer motion. This must be compensated by specific measures, which form part of the producer's know-how and usually are not published. The total thickness variation is typically below 2 µm for 150 mm wafers. Owing to the larger characteristic dimensions of GaAs-based devices, the demands for local flatness and particle content are lower than for silicon.

Particles on the surface of Si GaAs wafers, characterized by the density of light point defects (LPDs) are of two different origins: "classical" particles from the surroundings, which exhibit an isotropic scattering behaviour under a laser beam, and crystal-originated particles (COPs), which show anisotropic scattering. COPs are caused by As precipitates, which form pits or hillocks when they approach the surface owing to differences in the chemical potential of arsenic between the precipitates and the matrix. As the size of the As precipitates increases with decreasing dislocation density, the size distribution of the COPs is shifted to larger particles for VB/VGF-grown material, possibly conflicting with the customer's specification. Special measures are necessary to solve this problem.

An "epi-ready" surface quality of the wafers is obtained by a producer-specific processing step that leads to a mixed oxide layer with a specific composition, thickness and homogeneity, which can be removed by a desorption step prior to epitaxial growth. Furthermore, this surface layer, in conjunction with the packaging, prevents surface degradation during storage. As light-assisted chemical reactions can influence the stability of the "epi-ready" surface layer, packaging under nitrogen is common practice. As a consequence of the different "epi-ready" procedures of different wafer producers, the desorption of the protective layer will usually require modified desorption conditions. The "epi-ready" surface is exclusively of hydrophilic nature. The microroughness of GaAs wafers is characterized by RMS values in the range from 0.1 to 0.2 nm; the corresponding value for Si wafers is 0.1 nm.

21.6 The Future of GaAs and III–V Compounds

There is no doubt that semiconducting III–V compounds will not only hold but extend their present-day position in relation to optoelectronic devices. In addition to high-brightness diodes for illumination, covering the entire visible spectral range, white LEDs will continuously expand and will finally replace incandescent lamps. Lasers will remain a basic component of fibre communication networks and of data storage systems. A reduction of the numbers of types of substrate used for optoelectronic devices is expected: GaAs for

high-brightness LEDs in the range of red to amber, with layer structures of the type InGaAlP/GaAsP; GaN for green to blue, with GaN/GaAlN layer structures; and GaAs, InP and GaN for laser diodes.

Concerning microelectronics, high-power amplifiers and switches for wireless communications systems and other rf applications, as well as high-temperature devices, will need GaAs, InP and GaN and their “alloys” in future. Epitaxy will be the key device technology for microelectronic applications too, with m(**meta-morphic**)HEMTs, p(**pseudo-morphic**)HEMTs and HBTs on SI GaAs, HBTs on InP, and pHEMTs on GaN.

In contrast to GaAs and other III-V compounds such as GaP, InP and GaSb, which are readily available as single crystals and can be made into wafers “bulk” GaN is not available. Owing to its thermochemical properties, GaN (and other group III nitrides) cannot be grown from the melt. Other methods are being studied, such as solution growth from nitrogen-saturated Ga or Ga–Na solutions under high pressure at high temperatures (e.g. [78]), ammonothermal synthesis (e.g. [79]) and sublimation growth. But despite steady progress, bulk crystals are not yet being produced in the quantities and sizes needed ($< 1 \text{ cm}^3$) to establish a cost-effective commercialization of GaN technology on this basis. Instead, the long-lasting lack of native GaN substrates has shifted growth efforts to heteroepitaxy on a variety of substrates such as sapphire (Al_2O_3), 6H-SiC, Si, GaAs, ZnO and LiAlO_2 , of which sapphire and to a lesser extent SiC are mostly used. By use of such substrates combined with epitaxial lateral overgrowth and other “tricks”, the density of defects (especially dislocations) has been significantly reduced, and although it has not yet reached that of solution-grown crystals it has finally enabled commercialization of GaN-based devices. As the currently most advanced stage of this development, the first “free-standing” GaN wafers grown by HVPE (**hydride vapour phase epitaxy**) on sapphire templates have appeared recently [80, 81].

References

1. A. Seidl: 50 Jahre III-V-Halbleiter: ein Blick in die Originalliteratur. dgkk-Mitteilungsblatt **75**, 19 (2002)
2. K. Jacobs: Einige Ergänzungen zur Geschichte der III-V-Halbleiter. dgkk-Mitteilungsblatt **76**, 17 (2002)
3. R. Gremmelmaier: Herstellung von InAs- und GaAs-Einkristallen. Z. Naturfor. **11a**, 511 (1956)
4. A. Steinemann, U. Zimmerli: Dislocation-free gallium arsenide single crystals. In: H.S. Peiser (ed.): *Crystal Growth* (Pergamon Press, Oxford 1967) pp. 81–87
5. J.P. Mullin, B.W. Straughan, W.S. Brickell: Liquid encapsulation crystal pulling at high pressure. Phys. Chem. Solids **26**, 782 (1965)
6. E.P.A. Metz, R.C. Miller, R. Mazelsky: A technique for pulling single crystals of volatile materials. J. Appl. Phys. **33**, 2016 (1962)

7. A. Azuma: Method and device for producing compound single crystal with high dissociation. Japan. Patent 60-11299 (1983)
8. H. Kohda, K. Yamada, H. Nakanishi, T. Kobayashi, J. Osaka, K. Hoshikawa: Crystal growth of completely dislocation-free and striation-free GaAs. *J. Cryst. Growth* **71**, 813 (1985)
9. N.S. Beljackaja, S.P. Grishina: Study of the dislocation structure in GaAs single crystals (in Russian). *Izvest. AN Neorg. Mater.* **3**, 1347 (1967)
10. W.A. Gault, E.M. Monberg, J.E. Clemans: A novel application of the vertical gradient freeze method to the growth of high quality III-V crystals. *J. Cryst. Growth* **74**, 491 (1986)
11. R. Hull (ed.): *Properties of Crystalline Silicon*, EMIS Datareviews Series No. 20, (INSPEC IEE, London 1999)
12. M.R. Brozel, G.E. Stillman (eds.): *Properties of Gallium Arsenide*, Emis Datareview Series No. 16, (INSPEC IEE, London 1996)
13. T.P. Pearsall (ed.): *Properties, processing and applications of Indium Phoshide*, Emis Datareviews Series No. 21, (INSPEC IEE, London 2000)
14. O. Madelung (ed.): *Semiconductors – Group IV elements and III-V compounds* (Springer, Berlin, Heidelberg, New York 1991)
15. J.H. Edgar, S. Strite, I. Akasaki, H. Amano, C. Wetzel (eds.): *Properties, processing and applications of gallium nitride and related semiconductors*, Emis Datareviews Series No. 23, (INSPEC IEE 1999)
16. Landolt-Börnstein. In: *Semiconductors*, ed. by O. Madelung, M. Schulz, H. Weiss (Springer-Verlag, Berlin, 1984) pp. 12–34
17. J. Wu, W. Walukiewicz, K.M. Yu, J.W. Ager III, H. Lu, W.J. Schaff, Y. Saito, Y. Nanishi: Unusual properties of the fundamental band gap of InN. *Appl. Phys. Lett.* **80**, 3967–3969 (2003)
18. S.M. Sze: *High Speed Semiconductor Devices* (John Wiley & Sons, New York Chichester Brisbane Toronto Singapore 1990), p. 543
19. <http://www.ioffe.rssi.ru/SVA/NSM>
20. A.S. Jordan, A.R. von Neida, R. Caruso: The theoretical and experimental fundamentals of decreasing dislocations in melt grown GaAs and InP. *J. Crystal Growth* **76**, 243–262 (1986)
21. G.O. Meduoye, D.J. Bacon, K.E. Evans: Computer modelling of temperature and stress distribution in LEC-grown GaAs crystals. *J. Crystal Growth* **108**, 627–636 (1991)
22. A.S. Jordan: Some thermal and mechanical properties of InP essential to crystal growth modeling. *J. Crystal Growth* **71**, 559–565 (1985)
23. H. Gottschalk, G. Patzer, H. Alexander: Stacking fault energy and ionicity of cubic III V compounds. *phys. stat. sol. (a)* **45**, 207–217 (1978)
24. F. Berger, U. Bergmann, M. Schaper, R. Hammer, M. Jurisch: Microhardness testing of GaAs single crystals. *Materialprüfung* **4**, 117–122 (2001)
25. K. Yasutake, Y. Konishi, K. Adachi, K. Yoshii, M. Umeno, H. Kawabe: Fracture of GaAs wafers. *Jpn. J. Appl. Phys.* **27**, 2238–2246 (1988)
26. I. Yoneanga, T. Hoshi, A. Usui: High temperature hardness of bulk single crystal GaN. *MRS Internet J. Nitride Semicond. Res* **5S1**, W3.90 (2000)
27. M. Tatsumi, K. Fujita: Melt growth of GaAs single crystals. In: *Recent Development of Bulk Crystal Growth*, ed. by M. Isshiki (Research Signpost, Trivandrum, India, 1998) pp. 47–95

28. R.E. Kremer, R. Bindemann, S. Teichert: GaAs substrate production: optimization for cost and device type. In: Compound Semiconductor Manufacturing Expo, November 11–13, 2002, San Jose, CA
29. R. Bindemann: Volume production of GaAs substrates optimized with respect to characteristics and costs of device type and process technology. In: Gorham Conference Europe, June 17–19, 2002, Copenhagen
30. H. Wenzl, W.A. Oates, K. Mika: Defect thermodynamics and phase diagrams in compound crystal growth processes. In: D.T.J. Hurle (ed.): *Handbook of Crystal Growth*, vol. 1A (North-Holland, Amsterdam 1993) pp. 103–186
31. J. Korb, T. Flade, M. Jurisch, A. Köhler, T. Reinhold, B. Weinert: Carbon, oxygen, boron, hydrogen and nitrogen in the LEC growth of SI GaAs: a thermochemical approach. *J. Cryst. Growth* **198/199**, 343 (1999)
32. C.G. Kirkpatrick, R.T. Chen, D.E. Holmes, P.M. Asbeck, K.R. Elliott, R.D. Fairman, J.R. Oliver: LEC GaAs for integrated circuit application. In: R.K. Willardson, A.C. Beer (eds.): *Semiconductors and Semimetals*, Vol. 20 (Academic Press, New York 1984) pp. 159–231
33. D.T.J. Hurle: A mechanism for twin formation during Czochralski and encapsulated vertical Bridgman growth of III–V compound semiconductor. *J. Cryst. Growth* **147**, 239 (1995)
34. K. Terashima: Control of growth parameters for obtaining highly uniform large diameter LEC GaAs. In: *Semi-Insulating III–V Materials*, ed. by G. Grossmann, L. Ledebro (IOP Publishing, Bristol 1988) pp. 413–422
35. E. Molva, P. Bunod, A. Chabli, A. Lombardot, S. Dubois, F. Bertin: Origin of microscopic inhomogeneities in bulk gallium arsenide. *J. Cryst. Growth* **103**, 91 (1990)
36. O. Oda, H. Yamamoto, K. Kainosh, T. Imaizumi, H. Okazak: Recent developments of bulk III–V materials: annealing and defect control. In: J. Jimenez (ed.): *Defect Recognition and Image Processing in Semiconductors and Devices*, Vol. 135 (IOP Publishing, Bristol 1994) pp. 285–294
37. B. Hoffmann: Ein Beitrag zur thermischen Nachbehandlung von semiisokierenden LEC-GaAs-Einkristallen und -Scheiben. PhD Thesis, Universität Erlangen-Nürnberg (1996)
38. T. Steinegger: Defect engineering. PhD Thesis, TU Bergakademie Freiberg (2001)
39. F. Wirbeleit: The atomic structure of point defects in semiconductors by improved EPR/ENDIR data analysis. PhD Thesis, TU Bergakademie Freiberg (1998)
40. D.E. Holmes, R.T. Chen, K.R. Elliott, C.G. Kirkpatrick: Stoichiometry-controlled compensation in liquid encapsulated Czochralski GaAs. *Appl. Phys. Lett.* **40**, 46 (1982)
41. P.M. Petroff, L.C. Kimerling: Dislocation climb model in compound semiconductors with zinc blende structures. *Appl. Phys. Lett.* **29**, 461 (1976)
42. W.A. Oates, H. Wenzl: Foreign atom thermodynamics in liquid gallium arsenide. *J. Cryst. Growth* **191**, 303 (1998)
43. <http://gttserv.lth.rwth-aachen.de/gtt/>
44. M. Jurisch, D. Behr, R. Bindemann, T. Bünger, T. Flade, W. Fliegel, R. Hammer, S. Hölzig, A. Kiesel, A. Kleinwechter, A. Köhler, U. Kretzer, A. Seidl, B. Weinert: State-of-the-art semi-insulating GaAs substrates. In: K.H. Ploog, G. Tränkle, G. Weimann (eds.): *Compound Semiconductors 1999*, Vol. 166 (IOP Publishing, Bristol 2000) pp. 13–22

45. T. Bünger, J. Stenzenberger, F. Börner, U. Kretzer, S. Eichler, M. Jurisch, R. Bindemann, B. Weinert, S. Teichert, T. Flade: Active carbon control during the VGF growth of semiinsulating GaAs. In: *GaAs ManTech2003*, May 19–22, 2003, Scottsdale, AZ
46. S. Eichler, A. Seidl, F. Börner, U. Kretzer, B. Weinert: A combined carbon and oxygen segregation model for the LCE growth of SI GaAs. *J. Cryst. Growth* **247**, 69 (2003)
47. J. Völkl: Stress in the cooling crystal. In: D.T.J. Hurle (ed.): *Handbook of Crystal Growth*, Vol. 2B (North-Holland, Amsterdam 1994) pp. 821–874
48. <http://www.wafertech.co.uk>
49. P. Bennema: Growth and morphology of crystals. Integration of theories of roughening and Hartman–Perdok theory. In: D.T.J. Hurle (ed.): *Handbook of Crystal Growth*, Vol. 1A (North-Holland, Amsterdam 1993) pp. 477–581
50. A. Anwar: GaAs Bulk & Epitaxial Wafers Markets & Trends. Strategy Analytics, www.strategyanalytics.com (2002)
51. D.T.J. Hurle, B. Cockayne: Czochralski growth. In: D.T.J. Hurle (ed.): *Handbook of Crystal Growth*, Vol. 2A (North-Holland, Amsterdam 1994) pp. 99–211
52. T. Flade, M. Jurisch, A. Kleinwechter, A. Köhler, U. Kretzer, J. Prause, T. Reinhold, B. Weinert: State of the art 6" SI GaAs wafers made of conventionally grown LEC-crystals. *J. Cryst. Growth* **198/199**, 336 (1999)
53. A. Seidl, S. Eichler, T. Flade, M. Jurisch, A. Köhler, U. Kretzer, B. Weinert: 200mm GaAs crystal growth by the temperature gradient controlled LEC method. *J. Cryst. Growth* **225**, 561 (2001)
54. T. Inada, S. Komata, M. Ohnishi, M. Wachi, H. Akiyama, Y. Otoki: Development of mass production line for 150 mm GaAs wafers. In: *GaAs ManTech1999*, April 19–22, 1999, Vancouver, British Columbia, Canada, pp. 205–208
55. M. Shibata, T. Inada, S. Kuma: Huge single crystals of GaAs grown by LEC method. *Denshi Tokyo* **32**, 119 (1993)
56. M. Neubert, P. Rudolph: Growth of semi-insulating GaAs crystals in low temperature gradients by using the vapour pressure controlled Czochralski method (VCz). *Progress in Crystal Growth and Characterization of Materials* **43**, 119 (2001)
57. K. Kohiro, K. Kainosh, O. Oda: Growth of low dislocation density InP single crystals by the phosphorus vapor controlled LEC method. *J. Electron. Mater.* **20**, 1013 (1991)
58. <http://www.axt.com>
59. H.C. Rampsperger, E.H. Melvin: The preparation of large single crystals. *J. Opt. Soc. Am.* **15**, 359 (1927)
60. K. Sonnenberg, E. Küssel: Developments in vertical Bridgman growth of large diameter GaAs. *III–Vs Rev.* **10**, 30 (1997)
61. J.E. Clemans, E.M. Monberg: Crystal growth method. US Patent 4,923,561 (1990)
62. E. Monberg: Bridgman and related growth techniques. In: D.T.J. Hurle (ed.): *Handbook of Crystal Growth*, Vol. 2A (North-Holland, Amsterdam 1994) pp. 51–97
63. J. Stenzenberger, T. Bünger, F. Börner, S. Eichler, T. Flade, R. Hammer, M. Jurisch, U. Kretzer, S. Teichert, B. Weinert: Growth and characterization of 200mm SI GaAs single crystals grown by the VGF method. *J. Cryst. Growth* **250**, 57 (2003)

64. H. Yamamoto, O. Oda, M. Seiwa, M. Taniguchi, H. Nakata, M. Ejima: Microscopic defects in Semi-insulating GaAs and their effect on the FET device performance. *J. Electrochem. Soc.* **136**, 3089 (1989)
65. R. Stibal, J. Windscheif, W. Jantz: Contactless evaluation of semi-insulating GaAs wafer resistivity using the time-dependent charge measurement. *Semicond. Sci. Technol.* **6**, 995 (1991)
66. W. Siegel, G. Kühnel, C. Reichel, M. Jurisch, B. Hoffmann: High-resolution resistivity mapping of bulk semi-insulating GaAs by point-contact technique. *Mater. Sci. Eng. B* **44**, 238 (1997)
67. H.D. Geiler, M. Wagner, H. Karge, M. Paulsen, R. Schmolke: Photoelastic stress evaluation and defect monitoring in 300-mm-wafer manufacturing. *Mater. Sci. in Semiconductor Processing* **5**, 445 (2003)
68. M. Schaper, M. Jurisch, H.J. Klauß, H. Balke, F. Bergner, R. Hammer, M. Winkler: Fracture strength of GaAs wafers. In: B. Michel, T. Winkler, M. Werner, H. Fecht (eds.): *Proceedings Micro Materials* (Verlag ddp goldenbogen, Dresden 2000) pp. 570–573
69. S. Kuma, Y. Otoki: Usefulness of light scattering tomography for GaAs industry. In: J. Jimenez (ed): *Defect Recognition and Image Processing in Semiconductors and Devices* Vol. 135 (IOP Publishing, Bristol 1994) pp. 117–126
70. H.C. Alt, Y. Gomeniuk, U. Kretzer: Far-infrared spectroscopy of shallow acceptors in semi-insulating GaAs: evidence for defect interactions with EL2. *Phys. Stat. Sol. (b)* **235**, 58 (2002)
71. W. Siegel, S. Schulte, G. Kühnel, J. Monecke: Hall mobility lowering in undoped n-type bulk GaAs due to cellular-structure related nonuniformities. *J. Appl. Phys.* **81**, 3155 (1997)
72. P.D. Green: Growth of GaAs ingots with high free electron concentrations. *J. Cryst. Growth* **50**, 612 (1980)
73. R.C. Newman: The lattice location of silicon impurities in GaAs: effects due to stoichiometry, the Fermi energy, the solubility limit and DX behaviour. *Semicond. Sci. Technol.* **9**, 1749 (1994)
74. T. Kawase, H. Yoshida, T. Sakurada, Y. Hagi, K. Kaminaka, H. Miyajima, S. Kawarabayashi, N. Toyoda, M. Kiyama, S. Sawada, R. Nakai: Properties of 6-inch semi-insulating GaAs substrates manufactured by vertical boat method. In: *GaAs ManTech1999*, April 19–22, 1999, Vancouver, British Columbia, Canada, pp. 125–128
75. F. Duderstadt: Anwendung der von Karman'schen Plättentheorie und der Hertz'schen Pressung für die Spannungsanalyse zur Biegung von GaAs Wafern im modifizierten Doppelringtest. PhD Thesis, Technische Universität Berlin (2003)
76. A. Kleinwechter, T. Bünger, T. Flade, M. Jurisch, A. Köhler, U. Kretzer, A. Seidl, B. Weinert: Mass production of large-size GaAs wafers at FREIBERGER. In: *GaAs ManTech2001*, May 21–24, 2001, Las Vegas, NV, pp. 57–60
77. R. Hammer: Mikrorissbildung im GaAs und deren Nutzung beim Draht- und Innenloch trennen in der Halbleiterwafer-Herstellung. PhD Thesis, Universität Erlangen-Nürnberg (2002)
78. S. Porowski: High pressure growth of GaN – new prospects for blue lasers. *J. Cryst. Growth* **166**, 583 (1996)
79. R.K. Douglas, J.W. Kolis: Crystal growth of gallium nitride in supercritical ammonia. *J. Cryst. Growth* **222**, 431 (2001)

80. http://www.sei.co.jp/news_e/press/02/02_16.html
81. Y. Oshima, T. Eri, M. Shibata, H. Sunakawa, K. Kobayashi, T. Ichihashi, A. Usui: Preparation of freestanding GaN wafers by hydride vapor phase epitaxy with void-assisted separation. *Jpn. J. Appl. Phys.* **42**, L1 (2003)

Part IX

New, Exciting Fields: Do They Amalgamate with Silicon?

22 Quantum Computation by Electron Spin in SiGe Heterostructures

F.A. Baron, K.L. Wang

22.1 Introduction

In this paper we review recent efforts in the area of the solid-state implementations of the quantum computer (QC) using the electron spin as the quantum memory unit (the qubit). Quantum computing is viewed as a strong alternative to conventional solid-state electronic devices, which will face serious problems in the very near future owing to the ultimate physical limits of the CMOS technology. We discuss the expected performance of the QC, along with the physical requirements to be satisfied and the experimental challenges to be overcome for the practical realization of the QC. We focus mainly on the SiGe QC and describe the operational principles such as the writing procedure by g -factor engineering, the control of entanglement of qubits, and the readout of the quantum information. Particular, we present some recent theoretical calculations for the electron g -factor in SiGe using the $\mathbf{k} \cdot \mathbf{p}$ perturbation theory, which show the possibility of successful g -factor engineering in Ge-rich SiGe heterostructures. Finally, upon comparison with the other solid-state QC proposals using pure Si, GaAs, and CdTe quantum dots, we conclude that currently SiGe has the strongest potential for actual implementation.

Since the early 1960s, the silicon CMOS technology has undergone a tremendous development towards smaller device feature sizes, which has allowed packing-density increases and device performance improvements such as higher switching speed and lower power dissipation. Figure 22.1 presents the number of electrons in MOS channels, taken from previous reports [1] and extrapolated from the ITRS Roadmap. It shows an extrapolation down to just a few electrons per device that will take place in the very near future. Intuitively, one can anticipate that a great many problems will be encountered when the single-electron limit is approached, such as poor performance stability, noise, low speed due to the parasitic capacitance of the interconnects, and actual electron loss during transport due to interfaces and other defects. As the device dimensions should also decrease with the number of electrons, sooner or later the Bohr radius of an electron localized near a donor in Si (which for the phosphorus donor is 2.5 nm in Si and 6.5 nm in Ge) will be comparable to the device channel length, making the source-drain tunneling leakage current a major problem in the conventional CMOS technology. A lot

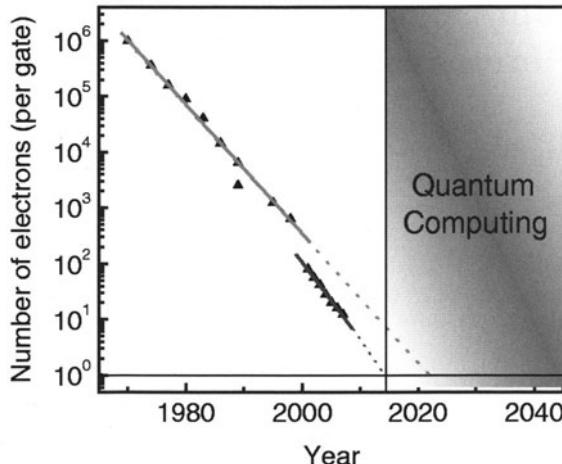


Fig. 22.1. International Technology Roadmap for Semiconductors showing the scaling trend for the number of electrons in a device channel extrapolated to a single electron per device. The figure serves as a motivation for research in the area of single-electron devices such as those using quantum bits (qubits). The data are taken from [1] and from anticipated values given by ITRS

of effort would have to be applied towards maintaining a very tight control of the electrostatic potential across the device, since the electron tunneling time and the tunneling current both have an exponential dependence on the barrier height and width. In order to fulfill such a task, one would need to maintain a precise quasi-atomic control over the source and drain doping profiles, a challenging task for the present technology exploiting ion implantation and lateral thermal diffusion. Since both of those processes produce a statistical distribution of the doping, one cannot ensure a uniform, abrupt doping profile.

We believe that the issues just mentioned indicate that the significant increase in the complexity of the technology is not an incidental effect but is, rather, a consequence of desperate attempts to implement the classical-physics-governed CMOS devices of a mesoscopic scale and to treat the legitimate quantum phenomena (such as tunneling) as parasitic effects, which is like squeezing a golf course into a residential area and ignoring peoples' complaints. It is the whole idea of *transistor* and *transport* that needs a serious reconsideration and, perhaps, abandonment in favor of quantum computing devices that operate consistently with their inherent nature, which is governed by the laws of quantum mechanics; that will reestablish the harmony between the simplicity of a device and high performance in the execution of complex tasks. A quantum circuit should have initial and final states and not be observed during the transition period. The computation is to be performed "blindly" by manipulating the overlap of the wave functions (entanglement)

of adjacent quantum bits (qubits), thereby excluding any problems of distant transfer and interconnects. Another advantage expected from the quantum devices is their dissipation-free operation, since the Schrödinger equation is dissipation-free, i.e. no potential can be attributed to the “friction” force. Thus, we are quite motivated to implement the concept of quantum computation and create the quantum computer.

22.2 The Expected Performance and Device Physics Issues of the QC

The idea of the “quantum mechanical computer” was first proposed by Feynman [2] and later gave rise to the concept of the “quantum Turing machine” and the quantum algorithms developed by Deutsch, Bernstein, Vazirani, Jozsa, Berthiaume, Brassard, Shor, Simon, and Grover (see the references in the review paper of Ekert and Jozsa [3]). The quantum computer consists of two-level quantum mechanical systems, which serve as the quantum bits (qubits), and its distinguishing feature is the ability to utilize coherent superpositions between the allowed physical energy states and to control the entanglement of the wave functions between two arbitrary qubits. From Shor’s paper [4], it follows that the Hilbert space of N qubits has at least $2^{2^N} - 1$ memory units owing to the capability of entanglement of all qubits’ wave functions. Having access to even one-tenth of a percent of such a swarm of entangled states would allow a simple 1000 qubit computer to possess a memory space equal to millions of today’s computers working together. In addition to that, the quantum parallelism phenomenon [3] enables an exponential computational speed-up for the problem of factorization into primes [4], as well as a general improvement in the computational efficiency for some other problems that are outlined elsewhere [5]. Even though quantum algorithms, in general, will not necessarily surpass their classical competitors in terms of their efficiency [6, 7], the quantum computer will offer superior computational power since its qubits, being separated by merely several Bohr radii, will have a much higher packing density.

Although one may certainly argue that a jump to quantum computing is quite a hard task for technology, the effort is justified because the performance of a QC below 20 nm should improve, while conventional CMOS tends to face more problems. However, there are still some serious obstacles to overcome on the road to the implementation of QC. Since the QC is an analog machine, errors due to dissipation (decoherence) through the interaction with the environment of the qubits are inevitable. Fortunately, people have invented efficient error correction codes to mitigate this problem. However, these codes effectively reduce the number of useful qubits, as many qubits are wasted for the purpose of backup. According to Preskill, about 125 qubits will be necessary to perform implementation of an effective error correction code

for each “computational” qubit [8]. Another requirement for fault-tolerant computation is that the decoherence time must be 10^4 times longer than the duration of a single computational cycle [9]. However, is quite possible that owing to the enormous capacity of Hilbert space and the enormous computational power, these issues will not become a deadlock for the creation of a QC.

From an engineering point of view, there are several requirements one needs to satisfy in order to cross the line between the observation of quantum effects and an actual device that utilizes them. Following the work of DiVincenzo et al. [10], we can summarize them as four basic requirements:

1. the ability to perform unitary transformations (rotations) of a single qubit,
2. accurately controlled entanglement of neighboring qubits,
3. a decoherence time exceeding 10^4 computational cycles,
4. projective measurements of qubits for the readout procedure.

22.3 SiGe Implementation of the QC Using Electron Spin

There are many two-level systems that may be exploited as qubits, but we shall focus on the approach involving the electron spin in SiGe, which has the best potential for solid-state implementation. An ability to manipulate the electron spin in SiGe would provide one with a robust two-level system suitable for quantum computing. Indeed, the choice of the SiGe system seems to be naturally friendly to quantum computing devices. From the technology side, it opens up the greatest opportunity to utilize the vast capabilities of the modern CMOS industry, such as superior-quality materials, scaling power, and easy integration with classical devices, which are still needed as a peripheral “wiring” to the quantum computer. Owing to the short spin decoherence time, an ultrafast operation speed must be maintained so that all computations will be finished before the quantum coherence is lost. So classical devices will be used to maintain high-frequency control over the quantum gates. It is certainly desirable that the frequency should not be too high, so the decoherence time must be as long as possible. From this perspective, silicon and germanium are also the best materials since their spin decoherence time T_2 is of the order of 10^{-3} sec [11–13]. For comparison, the spin decoherence time in bulk GaAs is only 10^{-7} sec [14].

Recently, we have proposed a quantum computer architecture which exploits silicon–germanium heterostructures, with a single electron spin performing as a qubit [15]. In that paper we suggested using phosphorus donor to nail down a single electron in a SiGe heterostructure; however, the replacement of phosphorus by SiGe quantum dots or simply by electrostatically formed quantum dots seems possible. We proposed to control the spin orientation using a gate-controlled spin resonance, which allowed us to meet

the first requirement on DiVincenzo et al.'s list. The gate exerts a local electric field on the nearby electron, forcing its wave function to shift across the SiGe layers with different Ge contents. As the g -factor is a function of the Ge content, its effective (average) value also changes with the gate potential. When the device is placed in global ac and dc magnetic fields, the electron g -factor entirely determines the spin resonance frequency, which follows from the resonance condition

$$\hbar\omega_R = \Delta E_{Zeeman} = -\mu_B \cdot \sum_{\alpha, \beta} g'_{\alpha\beta} \sigma_{\beta} B_{\alpha}, \quad (22.1)$$

where ω_R is the resonance frequency (typically in the microwave range), μ_B is the electron Bohr magneton, σ is the Pauli matrix, B is the magnetic field, and g_{ij} is the tensor of the g -factor (the g -tensor). Thus, the resonant frequency of an individual spin depends on the electric field applied and can be controlled by the gate. Suppose we fix the external radiation frequency by means of a cavity to some value with a small mismatch to the spin resonance frequency. By applying an appropriate voltage to the gate, we can have the electron g -factor match the condition (22.1) and expose the electron spin to the paramagnetic resonance in a controllable manner. This enables an arbitrary spin rotation to be obtained by simply applying an electric pulse to the gate for a certain time, equal to the appropriate fraction of the spin oscillation period.

The key assumption underlying this approach is that the electron g -factor is a sensitive function of the Ge content in the SiGe alloy, as suggested by the well-known difference between its values for bulk Si [16] ($g = 1.998$) and Ge [17] ($g = 1.57$). We are not aware of any systematic study of the electron g -factor in SiGe heterostructures, so we have carried out our own calculations for bulk SiGe based on the three-band $\mathbf{k}^* \mathbf{p}$ perturbation theory developed by Roth [13], taking account of strain and germanium content.

First, we realize that the conduction band minima in silicon and germanium are away from the Γ -point (in k -space) and therefore the symmetry is reduced to the symmetry of the group of the vector \mathbf{k}_0 which points to the conduction band minimum in k -space (the X-point for Si and L-point for Ge). Then, following Roth's paper, we take the electron g -factor to be a diagonal tensor:

$$g'_{\alpha\beta} = \begin{pmatrix} g_{xx} & 0 & 0 \\ 0 & g_{yy} & 0 \\ 0 & 0 & g_{zz} \end{pmatrix} = \begin{pmatrix} g_{\perp} & 0 & 0 \\ 0 & g_{\perp} & 0 \\ 0 & 0 & g_{||} \end{pmatrix}. \quad (22.2)$$

For the components of this tensor, Roth obtained the following expressions:

$$g_{zz} = 2 - \frac{\Delta_{so}}{E'_c - E'_v} \cdot \left(\frac{m}{m_{xx}} - 1 \right), \quad (22.3)$$

$$g_{xx} = 2 - \frac{\Delta_{so}}{E'_c - E'_v} \cdot \left(\frac{m}{m_{zz}} - 1 \right). \quad (22.4)$$

Here m_{zz} and m_{xx} are components of the electron effective-mass tensor, $E'_c - E'_v$ is the direct energy band gap at \mathbf{k}_0 , and Δ_{so} is the spin-orbit coupling interaction at \mathbf{k}_0 , which is estimated as 2/3 of the spin-orbit splitting of the valence band at the center of the Brillouin zone. The expressions (22.3) and (22.4) contain the energy band gap and the effective masses modified by the strain. For a SiGe heterostructure, the strain can be assumed to be a linear function of the difference between the Ge contents in the substrate and the epilayer, and thus the calculation of the g -tensor requires only a knowledge of the composition of the heterostructure, provided that the dependences of the energy band gap and of the effective mass on the strain are known. We can approximate the band gap as a linear function of the strain and deduce its value using a bilinear interpolation between the data [18] for the following structures: Si on Ge, bulk Si, Ge on Si, and bulk Ge. The effective mass can be obtained from theoretical data [19].

Note that the g -tensor given by (22.3) and (22.4) does not depend on the magnetic-field orientation or any other external parameters; its components are referred to a coordinate system where the z -axis coincides with the \mathbf{k}_0 vector, and the x - and y -axes are orthogonal to \mathbf{k}_0 and to each other. This is the principal coordinate system for the effective-mass tensor. Normally, the band edge is sixfold degenerate in bulk silicon and eightfold degenerate in germanium, so the effective g -factor (as measured in real space) will be an average over contributions from differently oriented ellipsoids. The strain will partially lift the degeneracy in silicon, causing redistribution of the electrons over the ellipsoids. Figure 22.2 illustrates how this works for the case of bi-

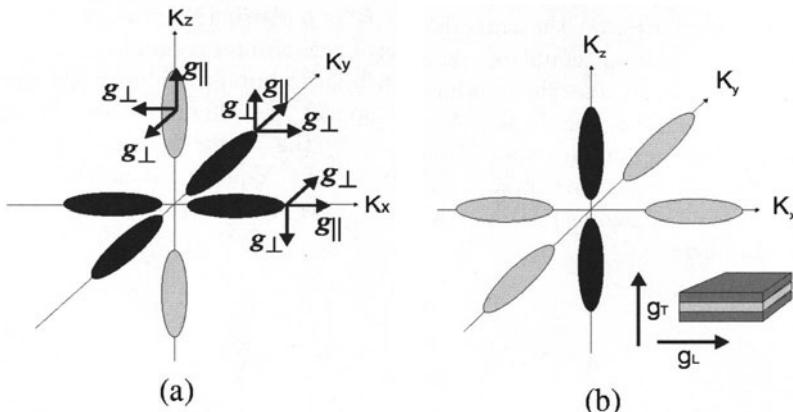


Fig. 22.2. Electron distribution over the constant-energy surfaces of the silicon conduction band minima in k -space: (a) Δ_{xy} minima (compressive strain) and (b) Δ_z minima (tensile strain). The orientation of the components of (a) the g -tensor and (b) the effective g -factor are shown as well

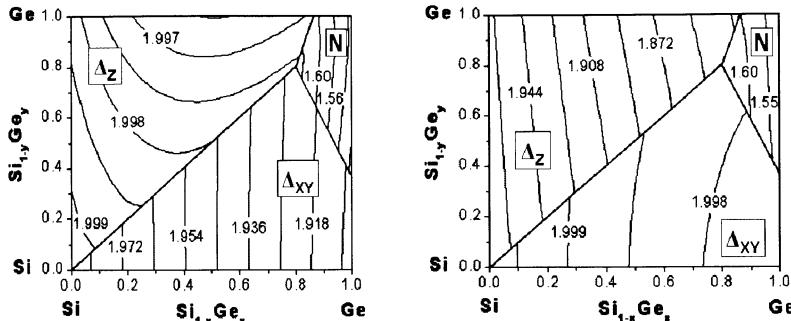


Fig. 22.3. Effective g -factor values for the components (a) g_L (longitudinal) and (b) g_T (transverse) in real space relative to the heterostructure interface plane

axial strain in the xy plane in silicon. It also shows the difference between the single-electron g -tensor and the components of the effective g -factor. Figure 22.3 gives the values of the effective g -factor components for all possible $\text{Si}_{1-x}\text{Ge}_x/\text{Si}_{1-y}\text{Ge}_y$ compositions at zero temperature. The x -axis represents the Ge content in the epilayer and the y -axis represents the Ge content in the substrate. Δ_z and Δ_{xy} correspond to the energy minima in the z direction and xy plane (Z and X valleys) caused by tensile and compressive strains, respectively. N stands for the range of Ge-like strained alloys with an energy minimum near the $\langle 111 \rangle$ direction. The coordinate system is chosen such that its z -axis is perpendicular to the interface between the two alloys. The result of our calculation shows that in order to obtain a large tuning range of the g -factor, one needs to use heterostructures with a high Ge content close to the transition between the X and L valleys, which is at approximately $x = 85\%$. An example of a quantum well containing two layers is shown in Fig. 22.4: the first layer one possesses a conduction band minimum in the X valley ($x < 85\%$), while the second layer has its minimum in the L valley ($x > 85\%$). According to our calculations (Fig. 22.3), the g -factor can be varied within a range of 1.5 to 1.9. However, we cannot make use of the whole range for tuning the g -factor since the L-X valley transition should be avoided, as it may not be fast enough unless it is phonon-assisted. Generally, we would like to avoid phonons, because they substantially enhance the decoherence rate [20] as the spin is coupled to lattice vibrations through the spin-orbit interaction. So we are forced to remain within the range of Ge-rich SiGe alloys with the conduction minimum in the L valley. According to Fig. 22.3, this will shrink the g -factor variation to the interval from 1.5 to 1.65, i.e. to a maximum change of 9%.

Figure 22.5 shows a schematic view of a cross section of a device and a qubit operation. When positively biased, the top gate shifts the electron wave function toward the higher-Ge-content layer, driving the spin into resonance and thereby implementing the unitary rotations described above. By apply-

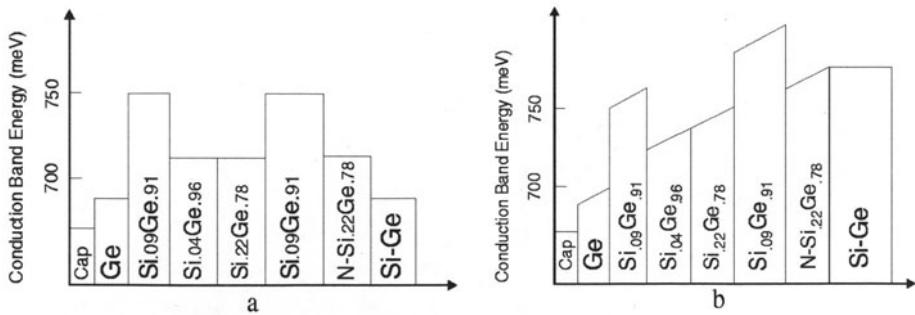


Fig. 22.4. Conduction band diagrams for SiGe qubit: (a) in the absence and (b) in the presence of an electric field. Layers with different SiGe composition have different work-function values, therefore quantum well is formed, where the g -factor can be engineered by the electric field

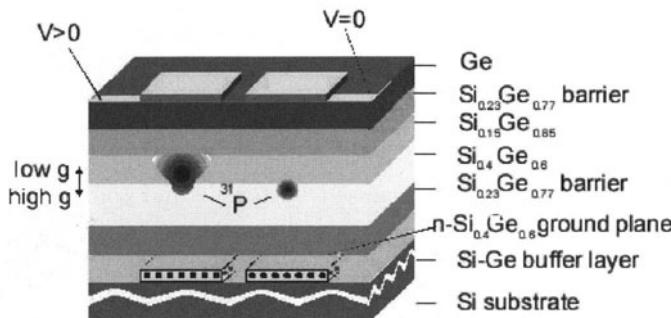


Fig. 22.5. A cross section of a device containing a qubit, showing the position of a phosphorus atom at the interface between two SiGe layers with a high and a low Ge content. Also shown is the effect of an attracting electric field on the wave function, resulting in its lateral diffusion due to the larger electron Bohr radius in the Ge-rich alloy

ing a higher positive bias, the electron can eventually be delocalized from the phosphorus donor and come into entanglement with a neighbor, as the Bohr radius of an electron occupying the L valley (Ge-like layer) is about three times larger than that of one occupying the X valley (Si-like layer). The entanglement can be triggered instantaneously and then terminated once the gate voltage has been set to zero. This fulfills the second requirement on DiVincenzo et al.'s list. The requirement for a sufficiently long decoherence time dictates a fast operational speed of the QC. Considering that the decoherence time in bulk germanium is 1 msec, the computational clock cycle should not exceed 10 nsec, and hence the operational frequency of the QC should be higher than 100 MHz. This should be easily achieved, since it is well below the conventional microprocessor speed of 2 GHz and so the third requirement on DiVincenzo et al.'s list is also satisfied.

Finally, according Kane [21], by measuring the conductivity of the channel placed underneath the phosphorus donor, one can implement a readout procedure. The channel conductivity depends on the charge state of the phosphorus atom. This charge state depends upon the final spin state of the electrons after entanglement: electrons in a singlet state will occupy one donor site, while a triplet state should force the electrons apart to minimize the energy of the system. Thus measuring the final charge state of the qubit allows determination of the relative orientation of the two spins. If one knows the initial state of one spin (the probe spin) then one should be able to determine the state of the other, which is the purpose of the readout procedure. The last requirement on DiVincenzo et al.'s list is thus satisfied.

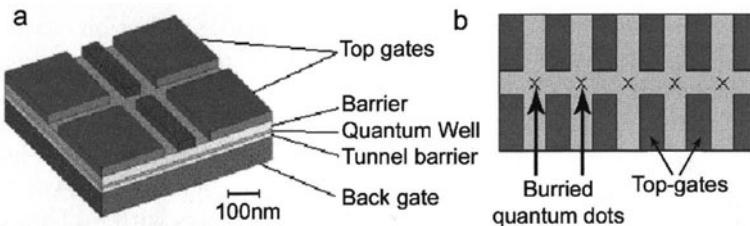


Fig. 22.6. Top and cross-sectional views of a quantum computer based on pure Si. Negatively biased top metal electrodes deplete the region of the silicon quantum well underneath, forming electrostatic quantum dots to host the single electrons that serve as the quantum bits. By changing the potential of the electrodes, one can produce entanglement between any adjacent quantum dots

22.4 Alternative Proposals

22.4.1 Pure Si Quantum Dots for QC

Another approach to the solid-state implementation of the QC, using pure Si quantum dots formed by electrostatic gates positioned on top of a Si quantum well (Fig. 22.6), has emerged recently [22]. People would like to give up the whole concept of *g*-factor engineering and spin resonance to simplify the QC device and its operation, by exploiting just the exchange interaction as suggested by DiVincenzo et al. [23]. The idea is to represent each qubit by a group of three spins, with the computational states coded as

$$|0\rangle = \sqrt{\frac{1}{2}} (|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle) \cdot |\uparrow\rangle, \quad (22.5)$$

$$|1\rangle = \sqrt{\frac{2}{3}} |\uparrow\uparrow\rangle |\uparrow\rangle - \sqrt{\frac{1}{6}} (|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle) \cdot |\uparrow\rangle, \quad (22.6)$$

and to precisely control the exchange interaction time integral

$$I_{j,j+1} = \int_0^\tau J(V_{gate}(t)) \cdot (s_j \cdot s_{j+1} + 1/2) dt \quad (22.7)$$

so that individual spins would rotate to a certain angle governed by the operator

$$U_{j,j+1} = \exp\left(\frac{i}{\hbar} I_{j,j+1}\right).$$

Thus, the coding (22.5) and (22.6) and an appropriately set exchange interaction between qubits (22.7) could serve as a nice remedy for the problem of rotation of a single qubit.

Despite its conceptual simplicity and the dramatic reduction of the complexity of the QC device, such an implementation remains challenging from the experimental point of view. According to [23], one has to maintain a time precision of 10^{-7} for the exchange coupling within a very short computational cycle. Since the clock cycle must be 10^5 times shorter than the decoherence time, then even for isotopically purified ^{28}Si with a coherence time of ~ 1 ms, one would have to control the switching time of the gates with ~ 1 fs precision in order to satisfy the requirement of 7-digit precision. The same accuracy must be attained in the magnitude of the exchange interaction. Failure to maintain the overlap integral (22.7) within the required precision would lead to computational errors arising from *under-* and *over*-exposure to the exchange coupling, which would result in imprecise spin rotations. Unlike the *g*-factor, the single-electron wave function and its overlap with other wave functions are not directly measurable entities, so experimentalists may end up struggling to calibrate the overlap integral versus the actual gate voltage for every single pair of qubits and may never achieve 7-digit precision. In a sense the overall complexity of the implementation of the QC is not actually reduced but is, rather, focused on the experimental part. Thus, the approach of single-spin manipulation seems more promising. Could one think of anything else but *g*-factor engineering to control a single spin?

22.4.2 GaAs and CdTe Quantum Dots for QC

In the SiGe proposal, the onset of spin resonance is caused by *g*-factor adjustment due to the relocation of the electron wave function across layers with different germanium contents. There is, however, another proposal [24], which suggests using GaAs or CdTe quantum dots and changing the spin-splitting energy (and thus tuning the resonance frequency) by applying a high electric field. The underlying physics is the Rashba effect [25,26], and we shall refer to this approach as the Rashba QC. This phenomenon is analogous to the effect in classical electrodynamics where an electric field transforms into an effective magnetic field in a moving frame of reference. This effective field lifts the spin

degeneracy. Since no external magnetic field is applied, the Rashba QC does not make use of *g*-factor engineering, but rather is a direct spin-splitting engineering by an electric field. Quantum Hall measurements on InGaAs/AlGaAs quantum wells have revealed a variation of the spin-splitting energy of 5–6 meV (10%) when an electric field of order 10⁵V/cm was applied [27], which was ascribed to the Rashba effect. This experiment, conducted on a 2D electron gas system, demonstrated that the absolute spin splitting was a linear function of the Fermi momentum \mathbf{k}_F $\Delta_R = E_\uparrow - E_\downarrow = 2\alpha \cdot k_F$, where α is a material-specific constant. However, in order to suppress decoherence, an actual QC will operate using single electrons confined in low-dimensional structures (such as quantum dots or quasi-quantum wires), where the electron momentum and Rashba splitting are likely to be very small. Despite having a favorable relative change of the spin splitting, its absolute magnitude will be too small to provide sufficient frequency bandwidth for tuning the spin resonance. Other issues of the Rashba QC include problems arising from an extremely low operational temperature and from short decoherence times in III–V compound materials, as was mentioned above.

Considering successful experiments [28, 29] on gate-controlled spin resonance in GaAs/AlGaAs quantum wells, which is believed to occur because of the *g*-factor change caused by the wave function relocation phenomenon, it is reasonable to suggest that the Rashba mechanism in SiGe will be also less efficient than *g*-factor engineering for spin control.

22.5 Conclusion

We have discussed the solid-state implementation of quantum computing as a natural extension of the CMOS technology into the nanoscale regime and have given an overview of current efforts in this area. In our opinion, SiGe heterostructures are the best candidates for hosting the electron spin qubits, while electron *g*-factor engineering represents the most practicable approach to the task of manipulation of single spins. As an initial guidance for the actual device design, we have calculated the electron *g*-factor in strained SiGe alloys using a formula emerging from the three-band $\mathbf{k} \cdot \mathbf{p}$ theory and by interpolation of the input parameters from the bulk Si and Ge values taken from the literature. The results imply that Ge-rich alloys (Ge > 85%) are a favorable choice. Overall, SiGe heterostructures seem to provide an optimal solution for the solid-state QC.

References

1. H. Iwai: Microelectron. J. **29**, 671 (1998); Semiconductor Industry Association, International Technology Roadmap for Semiconductors (ITRS). See website <http://public.itrs.net/>

2. R.P. Feynman: Found. Phys. **16**, No. 6 (1986)
3. A. Ekert, R. Jozsa: Rev. Mod. Phys. **68**, 733 (1996)
4. P.W. Shor: Phys. Rev. A **52**, R2493 (1995)
5. D.P. DiVincenzo, D. Loss: J. Magn. Magn. Mater. **200**, 202 (1999)
6. Y. Ozhigov: quant-ph/9712051
7. Y. Ozhigov: quant-ph/9803064
8. Private communications with Prof. J. Preskill
9. J. Preskill: Proc. R. Soc. London A **454**, 385 (1998)
10. D.P. DiVincenzo: cond-mat/9911245
11. M. Chiba, A. Hirai: J. Phys. Soc. Japan **33**, 730 (1972)
12. H. Hasegawa: Phys. Rev. **118**, 1523 (1960)
13. L.M. Roth: Phys. Rev. **118**, 1534 (1960)
14. W.H. Lau, J.T. Olesberg, M.E. Flatte: Phys. Rev. B **64**, 161301 (2001)
15. R. Vrijen, E. Yablonovitch, K.L. Wang, H.W. Jiang, A. Baladin, V. Roychowdhury, T. Mor, D.P. DiVincenzo: Phys. Rev. **62**, 012306 (2000)
16. H. Vollmer, D. Geist: Phys. Status Solidi B **62**, 367 (1974)
17. G. Feher, D.K. Wilson, E.A. Gere: Phys. Rev. Lett. **3**, 25 (1959)
18. Q. Ma, K.L. Wang, J.N. Schulman: Phys. Rev. B **47**, 1936 (1993); C. Tsferbakh, H.M. Polatoglou, G. Theodorou: Phys. Rev. B **47**, 7104 (1997); J.R. Chelikowsky, M.L. Cohen: Phys. Rev. B **14**, 556 (1976); *Landolt-Börnstein, Zahlenwerte und Funktionen aus Naturwissenschaften und Technik*, ed. by K.H. Hellwege, New Series, Group III, vol. 17a, ed. by O. Madelung, M. Schulz, H. Weiss (Springer, Berlin 1982); L.D. Laude, F.H. Pollak: Phys. Rev. B **3**, 2623 (1971); A.R. Goni, K. Syassen, M. Cardona: Phys. Rev. B **39**, 12921 (1989); I. Balslev: Phys. Rev. **143**, 636 (1966); *Solids Under Pressure*, ed. by W. Paul, D.M. Warschauer (McGraw-Hill, New York 1963); M. Chandrasekhar, F.H. Pollak: Phys. Rev. B **15**, 2127 (1977); J. Weber, M.I. Alonso: Phys. Rev. B **40**, 5683 (1989)
19. M. Rieger, P. Vogl: Phys. Rev. B. **48** (19), 14276 (1993)
20. D. Mozyrsky, Sh. Kogan, V.N. Gorshkov, G.P. Berman: Phys. Rev. B **65**, 245213 (2002)
21. B.E. Kane: Nature **393**, 133 (1998)
22. M. Friesen, P. Rugheimer, D.E. Savage, M.G. Lagally, D.W. Weide, R. Joynt, M.A. Eriksson: cond-mat/0204035
23. D.P. DiVincenzo, D. Bacon, J. Kempe, G. Burkard, K.B. Whaley: Nature **408**, 339 (2000)
24. S. Bandyopadhyay: Phys. Rev. B **61**, 13813 (2000)
25. Y.A. Bychkov, E.I. Rashba: Sov. Phys. JETP Lett. **39**, 78 (1984)
26. E.L. Ivchenko, A.A. Kiselev, M. Willander: Solid State Commun. **102**, 375 (1997)
27. J. Nitta, T. Akazaki, H. Takayanagi, T. Enoki: Phys. Rev. Lett. **78**, 1335 (1997)
28. H.W. Jiang, E. Yablonovitch: Phys. Rev. B **64**, 041307 (2001)
29. G. Salis, Y. Kato, K. Ensslin, D.C. Driscoll, A.C. Gossard, D.D. Awschalom: Nature **414**, 619 (2001)

23 Carbon Nanotube Applications in Microelectronics

W. Hoenlein, F. Kreupl, G.S. Duesberg, A.P. Graham, M. Liebau, R. Seidel, E. Unger

23.1 Introduction

The extraordinary characteristics of carbon nanotubes make them a promising candidate for applications in microelectronics. Catalyst-mediated CVD growth is very well suited for selective, *in situ* growth of nanotubes compatible with the requirements of microelectronics technology. This deposition method can be exploited for carbon nanotube vias. Semiconducting single-walled tubes can be successfully operated as carbon nanotube field effect transistors (CNTFETs). A simulation of an ideal CNTFET is presented and compared with the requirements of the ITRS Roadmap. Finally, we compare an upgraded CNTFET with the most advanced silicon MOSFETs.

Carbon nanotubes are a new modification of carbon discovered in 1991 by Iijima [1] while looking at soot residues from a fullerene experiment. However, unlike fullerenes, which are ball-type structures composed of carbon hexagons and pentagons, carbon nanotubes are long tubes with a purely hexagonal (graphitic) structure (Fig. 23.1).

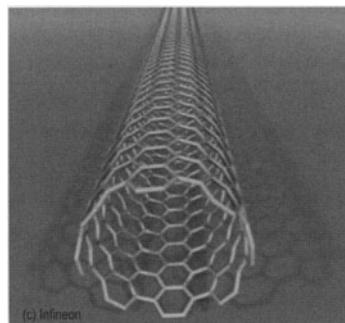


Fig. 23.1. Carbon nanotube structure. The tube can be considered as a seamlessly rolled-up graphite sheet

Indeed, carbon nanotubes can be considered as a rolled-up single layer of graphite, with diameters ranging roughly from 1 to 100 nm and up to millimeters in length. Owing to the extraordinary strength of the carbon–carbon bond, the small atomic diameter of the carbon atom, and the availability of free π -electrons in the graphitic configuration, nanotubes exhibit a number of

Table 23.1. Some important electrical and mechanical characteristics of carbon nanotubes

Electrical conductivity	Metallic or semiconducting
Electrical transport	Ballistic, no scattering
Energy gap (semicond.)	E_g (eV) $\approx 1/d$ (nm)
Maximum current density	$\sim 10^{10}$ A/cm ²
Maximum strain	0.11% @ 1 V
Thermal conductivity	6000 W/(K m)
Diameter	1–100 nm
Length	Up to millimeters
Gravimetric surface	> 1500 m ² /g
Young's modulus E	1000 GPa

remarkable electronic and mechanical characteristics, which are summarized in Table 23.1.

For applications in microelectronics, the most interesting features are the ballistic (scattering-free) and spin-conserving transport of electrons along the tubes, the ability to have metallic as well as semiconducting behavior, and access to the energy gap, which depends on the diameter of the tube. Another interesting issue for microelectronics is the unsurpassed thermal conductivity, which is roughly twice that of diamond. These and other striking features and applications of carbon nanotubes are reviewed in [2, 3]. Semiconducting and metallic configurations of nanotubes can be achieved by rolling them up differently. Figure 23.2 shows that the so-called armchair configuration always leads to metallic tubes, whereas the zigzag configuration is semiconducting or metallic. In addition, there is also a way of rolling up nanotubes with a chirality. It turns out, that 2/3 of all possible configurations show semiconducting behaviour.

In Fig. 23.3, calculations of the density of states for a metallic and a semiconducting nanotube are shown [4]. The integers n , m are the vector parameters describing the chirality of the tube and correspond to a rolling-up axis on the two-dimensional graphite sheet.

The electron energy levels are split into subbands owing to the restricted geometry perpendicular to the tube axis. It is interesting to note that, unlike the case in other two-dimensional electron gas systems, the wave function does not sense the irregularities at the edge of the device, giving rise to potential fluctuations and additional scattering, but represents a perfectly terminated system only “disturbed” by the contacts. Thus, for the metallic configuration, extremely high current densities of more than 10¹⁰ A/cm² have been reported [5].

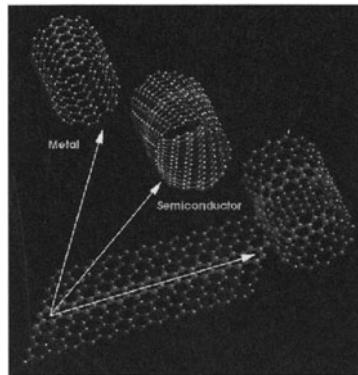


Fig. 23.2. Different morphological configurations and resulting electrical conduction types for carbon nanotubes [23]. *Left:* armchair configuration. *Right:* zigzag configuration. *Middle:* intermediate chirality

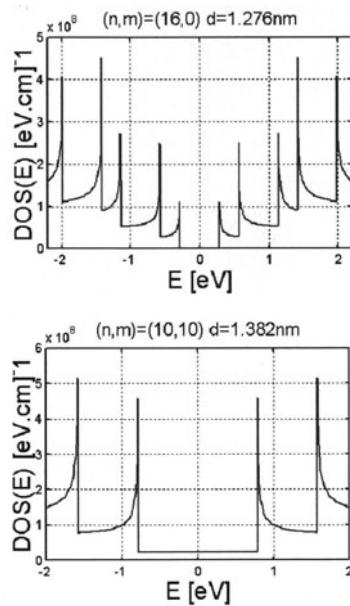


Fig. 23.3. Calculations of the density of states for a semiconducting (*top*) and a metallic (*bottom*) carbon nanotube. The tubes are characterized by the integers n , m , which determine their chirality

23.2 Nanotube Fabrication

A number of growth methods have been described in the literature, including an arc discharge between graphite electrode rods and laser-induced evaporation of graphite targets containing traces of various catalyst metals [6].

Nanotubes are then collected from the cold parts of the deposition system and dispersed in a solvent. Deposition of the nanotube-containing liquid and subsequent drying of the solvent leaves the nanotubes in randomly distributed places, some of them (hopefully) on contacts, where they can be used for electrical characterization. From the methods which are compatible with microelectronics, we selected CVD-based *in situ* growth as proposed by Dai [7]. Nanotubes were grown from a catalyst in a carbon-containing gas atmosphere at temperatures of 600–900°C. At the lower end of this temperature range the nanotubes grow with a number of concentric layers and are called multiwalled nanotubes, whereas at the higher end of the temperature band the nanotubes grow preferentially as single-walled species with small diameters. Microelectronic patterning methods can be used to structure the catalyst layer, thereby defining places for selective growth. Figure 23.4 shows the growth of nanotube blocks, each consisting of thousands of individual multiwalled nanotubes.

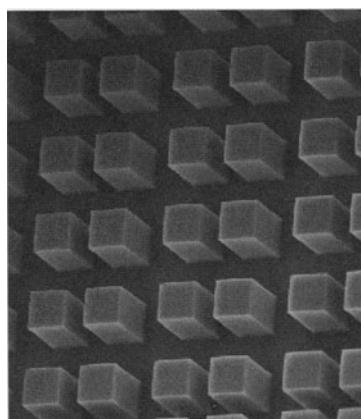


Fig. 23.4. Catalyst-mediated selective CVD growth of nanotube blocks (100 µm × 100 µm base) consisting of thousands of individual tubes

23.3 Carbon Nanotube Interconnects

The scattering-free transport of electrons possible in defect-free carbon nanotubes is a most attractive feature for microelectronic applications. Decreasing the thickness of conventional polycrystalline interconnects leads to additional scattering at the surfaces and grain boundaries, thus deteriorating the wire resistance [8]. Carbon nanotubes offer an undisturbed quasi-crystalline wire-like structure, where pulses can travel without being hindered by length-dependent ohmic scattering. In Fig. 23.5 we have estimated the signal delay with a very simple model, indicating that nanotubes could surpass classical wires with respect to signal delay, especially for long wires [9].

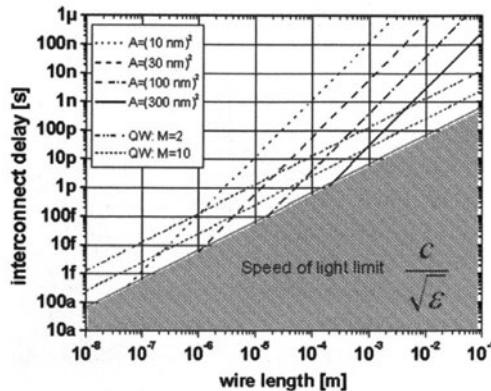


Fig. 23.5. Signal propagation in classical and quantum wires. Classical delays are governed by the RC time constants of wires, with different cross sections A . Quantum wires are governed by capacity only, with the conduction modes M as a parameter

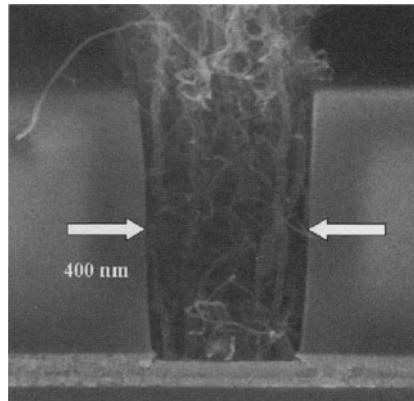


Fig. 23.6. Selective growth of multiwalled carbon nanotubes in a via. The top contact has not yet been formed

It should be emphasized, however, that the wire delay in a circuit is governed also by the interactions of wires and active devices. It should also be noted that contact resistances add to the quantum resistance which characterizes a carbon nanotube.

Combining the unsurpassed current density with the ability to grow nanotubes at specific sites, we developed a concept for a nanotube via. Vias are interconnects between wiring layers in chips and are prone to deterioration due to current crowding and subsequent electromigration. We propose to use carbon nanotubes instead of metal plugs to overcome this problem [9]. Figure 23.6 shows the selective growth of a great number of nanotubes in a via with state-of-the-art dimensions.

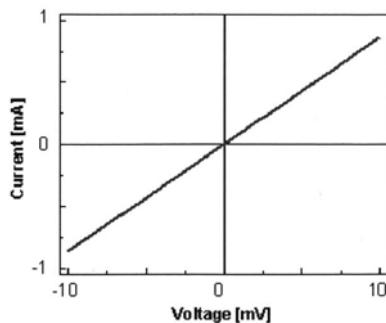


Fig. 23.7. Current–voltage characteristic of a carbon-nanotube-filled via showing ohmic behavior

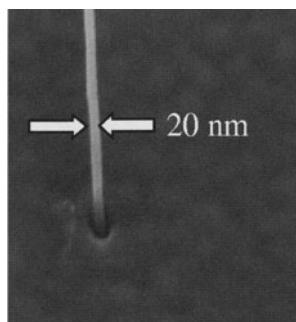


Fig. 23.8. Single multiwalled carbon nanotube growing out of a 20 nm hole

Figure 23.7 gives an example of an electrical characteristic of such a nanotube, indicating the presence of mostly metallic species described by a linear current–voltage relationship. The carbon nanotube via has the additional advantage of being scalable to nanometer dimensions with a large aspect ratio. As shown in Fig. 23.8, we have succeeded in growing one single multiwalled nanotube in a hole with a diameter of 20 nm [10]. It should be noted that the holes were fabricated with conventional lithographic methods using a spacer reduction method [11].

23.4 Carbon Nanotube Transistors and Circuits

Semiconducting single-walled carbon nanotubes were first demonstrated in 1998 to show a technologically exploitable switching behavior [12]. Tans et al. applied an electric field to the nanotube using a back gate and managed to modulate the conductivity over more than five orders of magnitude (Fig. 23.9).

Since this first demonstration of a carbon nanotube field effect transistor, tremendous progress has been made in improving the electrical characteristics

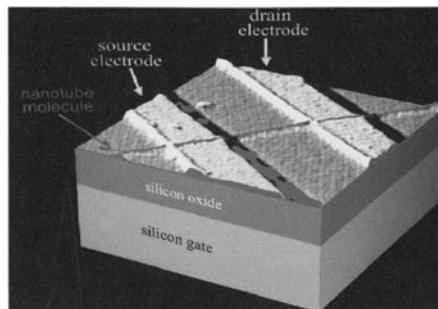


Fig. 23.9. First functional carbon nanotube field effect transistor (CNTFET) [12]. The nanotube is placed on top of two electrodes. A silicon dioxide layer on top of a silicon wafer acts as the back-gate dielectric

of the CNTFET. One disadvantage of Tans's arrangement was the back-gate dielectric, which consisted of a thick silicon dioxide layer. Bachthold et al. [13] replaced the silicon dioxide dielectric by a thin native Al_2O_3 layer on top of a patterned Al gate and were able to lower the gate voltage and to increase the transconductance. A further step was realized by the introduction of a top-gate device by Wind et al. [14]. The electric field at the nanotube was increased by improving the geometry of the gate electrode, and the contact resistance was reduced by a suitable choice of contact material, resulting in further-increased performance values. Additional progress was made by the creation of the first CMOS-like device, where a section of a nanotube was doped with potassium, thus reversing its conductivity type [15]. This device corresponds to a complementary NOR gate with a gain larger than unity. Other realizations of CNTFET-based circuits include simple SRAM cells and ring oscillators [13]. However, the performance of these devices is poor owing to associated off-chip wiring, nonoptimized device design, and the fact that only single nanotubes can be used as devices. Comparison with state-of-the-art silicon devices, however, requires one to assume virtually upgraded CNT devices, where an appropriate number of tubes are considered to work in parallel so as to correspond to the lateral dimensions of existing silicon devices. In the next section we shall perform some estimations for future CNT devices and make a comparison with the best silicon MOSFETs available.

23.5 CNTFET Simulations and Vertical-CNTFET Concept

The physical description of current transport in a CNTFET is still under debate. However, a recently proposed model [16] that assumes one-dimensional electrostatics fits quite well to existing experimental data [17]. It relates the charge in the tube to the capacitance of the CNTFET arrangement and the

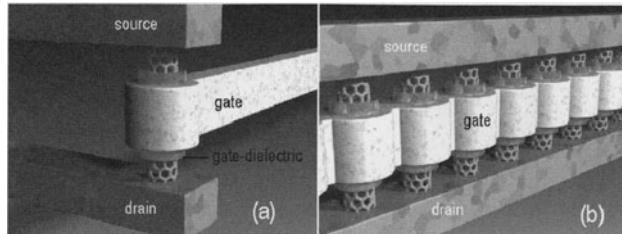


Fig. 23.10. The vertical-CNTFET (VCNTFET) concept (a). Parallel operation of many VCNTFETs (b)

applied gate voltage. On the basis of this theory one can estimate the best-case performance and compare it with the requirements for silicon devices for the year 2016, as defined by the ITRS Roadmap. We have proposed a vertical nanotube transistor [18] consisting of a 1 nm diameter, 10 nm long single-walled tube with a coaxial gate and a gate dielectric with a thickness equivalent to 1 nm of oxide (Fig. 23.10a).

In order to compare the proposal device with the corresponding silicon MOSFET, we assume the operation of 250 nanotubes per micron in parallel, as shown schematically in Fig. 23.10b. Using Guo et al.'s theory [16], we can estimate the performance and compare it with the respective Si MOSFET (Table 23.2). It can easily be deduced that the CNTFET outperforms the Si MOSFET by far.

Table 23.2. Comparison of the most important transistor characteristics for a year-2016 MOSFET (according to the ITRS Roadmap) and a VCNTFET

	V_{dd} (V)	Drive current ($\mu\text{A}/\mu\text{m}$)	Trans- conductance ($\mu\text{S}/\mu\text{m}$)	$t(C_{\text{gate}} \cdot S)$ (V_{dd}/I_{dd}) (ps)	S (mV/dec)	Leakage ($\mu\text{A}/\mu\text{m}$)	Effective t_{ox} (nm)
ITRS year 2016	0.4	1500	1000	0.15	70	10	0.4–0.5
CNT- FET	0.4	2500	15000	0.08	65	2.5	1

We can now ask what influence deviations from nonideal behavior have and whether they can be compensated by the performance surplus. Most probably, the k -vector mismatch at the metallic contacts (three-dimensional to one-dimensional) and other mechanisms will lead to a contact resistance, which we assume to be an ohmic series resistance (R_s , R_d) on both sides of the tube. Figure 23.11 shows the corresponding degradation of the transconductance, the drive current, and the output characteristics of the CNTFET as a function of the series resistance. If we assume a reasonable value of $50\text{ k}\Omega$ for each contact resistance, as indicated by the circles, there is still room for compliance with the ITRS values.

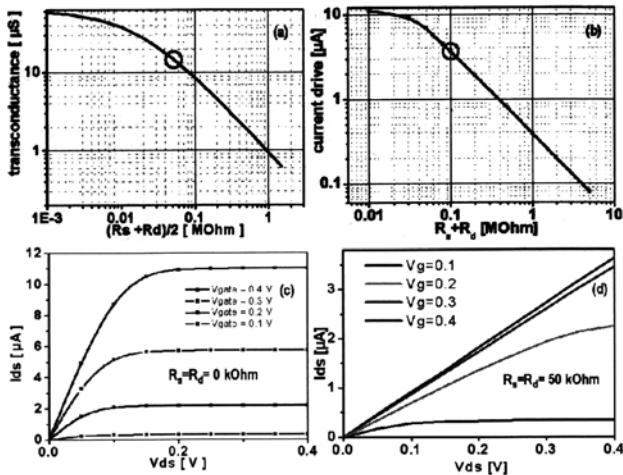


Fig. 23.11. Degradation of the ideal transconductance (a) and drive current (b) due to source and drain series contact resistances (Circles denote an assumed value of $R_s = R_d = 50 \text{ k}\Omega$). The ideal output characteristics (c) also degrade owing to contact series resistances (d)

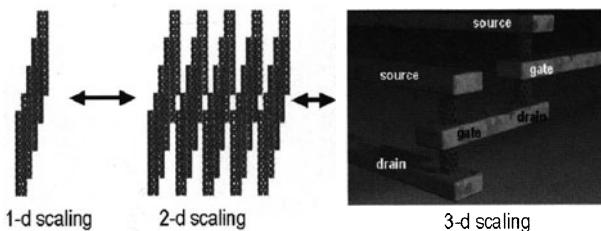


Fig. 23.12. Three-dimensional scaling by exploitation of the VCNTFET concept

The vertical-transistor concept also allows higher packing densities to be achieved, because source and drain areas can be arranged on top of each other. Furthermore, real three-dimensional structures are possible, because the active devices are no longer bound to the surface of the monocrystalline silicon wafer (Fig. 23.12).

The comparison of carbon-nanotube- and silicon-based transistors is not restricted to simulated devices for 2016, however, but can be performed for today's most advanced state-of-the-art CNTFETs and silicon FETs. In Table 23.3 we compare two p-type CNTFETs with three Si-MOSFETs, details of which were all published very recently. All data for the CNTFETs were measured on single-nanotube devices and have been normalized to device width for comparison. It can be seen that all values for the CNTFETs outperform those for the best MOSFET devices. It should be noted that Rosenblatt et al.'s CNTFET [19] is gated by an electrolyte and that nonstatic parameters

Table 23.3. Comparison of some important experimental transistor characteristics for CNTFETs and silicon MOSFETs

	p-CNTFET [19] 1.4 μ m (1 V) Rosenblatt et al. (2002)	p-CNTFET [17] 3 μ m (1.2 V) Javey et al. (2002)	MOSFET [20] 0.1 μ m (1.5 V) Ghani et al. (1999)	FinFET [21] 10 nm (1.2 V) Yu et al. (2002)	MOSFET [22] 14 nm (0.9 V) Doris et al. (2002)
Drive current I_{ds} (mA/ μ m)	2.99	3.5	1.04 nFET 0.46 pFET	0.450 nFET 0.360 pFET	0.215 pFET
Transconductance (μ S/ μ m)	6666	6000	1000 nFET 460 pFET	500 nFET 450 pFET	360 pFET
S (mV/dec)	80	70	90	125 101	71
On-resistance (Ω / μ m)	360	342	1442 nFET 3260 pFET	2653 nFET 3333 pFET	4186 pFET
Gate length (nm)	1400	2000	130	10	14
Normalized gate oxide (1/nm)	80/1 = 80	25/8 = 3.12	4/2 = 2	4/1.7 = 2.35	4/1.2 = 3.33
Mobility (cm ² /Vs)	1500	3000	—	—	—
I_{off} (nA/ μ m)	—	1	3	10	100

have not been included in the comparison. Current CNTFETs are still much longer than actual MOSFET channels but outperform silicon MOSFETs with respect to on-resistance, a key issue for RC delay. One of the most interesting questions of CNTFET development is whether short-tube devices will also surpass silicon devices in the frequency domain.

Since their first appearance in 1998, CNTFETs have made tremendous progress. It should be noted, however, that the parallel and dense formation of nanotube grids necessary for parallel production has not yet been realized. Moreover, carbon nanotube technology is in its infancy and is by no means comparable to the mature silicon technology. However, industry is now taking a strong interest and similar progress is anticipated on the technology side.

23.6 Conclusions

Carbon nanotubes offer some remarkable characteristics for microelectronics applications. Scattering-free current transport allows high current densities and improved signal delays for long wires. Selective catalyst-mediated CVD growth of carbon nanotubes is the most suitable method for carbon nanotube vias. CNTFETs with semiconducting nanotubes are promising candidates for transistor devices with superior characteristics compared with silicon MOSFETs. Vertical CNTFETs are proposed as key elements for a silicon-free three-dimensional integration scenario.

References

1. S. Iijima: Helical microtubules of graphitic carbon. *Nature* **354**, 56 (1991)
2. R.H. Baughman, A.A. Zakhidov, W.A. de Heer: Carbon nanotubes—the route toward applications. *Science* **297**, 787 (2002)
3. P.G. Collins, P. Avouris: Nanotubes for electronics. *Sci. Am.* **12**, 62 (2000)
4. J.W. Mintmire, C.T. White: Universal density of states for carbon nanotubes. *Phys. Rev. Lett.* **81** (12), 2506 (1998)
5. B.Q. Wei, R. Vajtai, P.M. Ajayan: Reliability and current carrying capacity of carbon nanotubes. *Appl. Phys. Lett.* **79** (8), 1172 (2001)
6. R. Saito, G. Dresselhaus, M.S. Dresselhaus: *Physical Properties of Carbon Nanotubes* (Imperial College Press, London 1998)
7. H. Dai: Carbon nanotubes: synthesis, integration, and properties. *Acc. Chem. Res.* **35**, 1035 (2002)
8. W. Steinhoegl, G. Schindler, G. Steinlesberger, M. Engelhardt: Size-dependent resistivity of metallic wires in the mesoscopic range. *Phys. Rev. B* **66**, 075414 (2002)
9. W. Hoenlein: New prospects for microelectronics: carbon nanotubes. *Jpn. J. Appl. Phys.* **41**, 4370 (2002)
10. D.S. Duesberg, A.P. Graham, M. Liebau, R. Seidel, E. Unger, F. Kreupl, W. Hoenlein: Growth of isolated carbon nanotubes with lithographically defined diameter and location. *Nano Lett.* (2003), <http://dx.doi.org/10.1021/nl025906c>
11. M. Engelhardt, G. Schindler, K. Mosig, G. Steinlesberger, W. Steinhoegl, G. Gebara: Extending copper metallization technology for wiring to end-of-the-roadmap feature sizes. Conference Proceedings Advanced Metallization Conference, Montreal (2001) pp. 11–17
12. S.J. Tans, A.R.M. Verschueren, C. Dekker: Room-temperature transistor based on a single carbon nanotube. *Nature* **393**, 49 (1998)
13. A. Bachthold, P. Hadley, T. Nakanishi, C. Dekker: Logic circuits with carbon nanotube transistors. *Science* **294**, 1317 (2001)
14. S.J. Wind, J. Appenzeller, R. Martel, V. Derycke, P. Avouris: Fabrication and electrical characterization of top gate single-wall carbon nanotube field-effect transistors. *Appl. Phys. Lett.* **80**, 3817 (2002)
15. V. Derycke, L. Martel, J. Appenzeller, P. Avouris: Carbon nanotube inter- and intramolecular gates. *Nano Lett.* **9**, 453 (2001)

16. J. Guo, M. Lundstrom, S. Datta: Performance projections for ballistic carbon nanotube field-effect transistors. *Appl. Phys. Lett.* **80**, 3192 (2002)
17. A. Javey, H. Kim, M. Brink, Q. Wang, A. Ural, J. Guo, P. McIntyre, P. McEuen, M. Lundstrom, H. Dai: High- K dielectrics for advanced carbon nanotube transistors and logic gates. *Nature Mater.* **1**, 241 (2002)
18. German Patent DE 0010036897 C1 (2000)
19. S. Rosenblatt, Y. Yaish, J. Park, J. Gore, V. Sazonova, P. McEuen: High performance electrolyte gated carbon nanotube transistors. *Nano Lett.* **2**, 869 (2002)
20. T. Ghani, S. Ahmed, P. Aminzadeh, J. Bielefeld, P. Charvat, C. Chu, M. Harper, P. Jacob, C. Jan, J. Kavalieros, C. Kenyon, R. Nagisetty, P. Packan, J. Sebastian, M. Taylor, J. Tsai, S. Tyagi, S. Yang, M. Bohr: 100 nm gate length high performance / low power CMOS transistor structure. *IEDM Technical Digest* (1999) pp. 415–419
21. B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C. Yang, C. Tabery, C. Ho, Q. Xiang, T. King, J. Bokor, C. Hu, M. Lin, D. Kyser: FinFET scaling to 10 nm gate length. *IEDM Technical Digest* (2002) pp. 251–254
22. B. Doris, M. Jeong, T. Kanarsky, Y. Zhang, R.A. Roy, O. Dokumaci, Z. Ren, F. Jamin, L. Shi, W. Natzle, H. Huang, J. Mezzapelle, A. Mocuta, S. Womack, M. Gibelyuk, E.C. Jones, R.J. Miller, H.P. Wong, W. Haensch: Extreme scaling with ultra-thin Si channel MOSFETs. *IEDM Technical Digest* (2002) pp. 267–270
23. <http://www.nas.nasa.gov/Groups/SciTech/nano/images/images.html>

24 Creating Systems for Ambient Intelligence

K. Delaney, J. Barton, S. Bellis, B. Majeed, T. Healy,
C. O'Mathuna, G. Crean

The future of information technology systems will be driven by the vision of ambient intelligence (AmI). In this vision, AmI will surround us with proactive interfaces supported by computing and networking-technology platforms that are everywhere; for instance, its systems would be embedded into everyday objects such as furniture, clothes, vehicles, roads, and even decorative materials such as paint and wallpaper. They would be unobtrusive, often invisible. They will provide a seamless environment of computing, advanced networking technology, and specific interfaces. The systems will be aware of the presence of human and of the characteristics of their personalities, and will take care of their needs. AmI will be capable of responding intelligently to spoken or gestured indications of desire, and could engage in intelligent dialogue. Interacting with AmI would be relaxing and enjoyable for the citizen and would not involve steep learning curves. The nature of this system will evolve from existing technologies, particularly silicon, as that is the primary platform upon which intelligence systems can be built. However, new approaches in semiconductor material and process development and systems in integration will be required to secure the future emergence of AmI.

24.1 Introduction

Ambient systems open up entirely new possibilities for future applications and resultant markets. Ultimately, these systems will create intelligent environments that cater continuously for the requirements of the individual in everyday life and apply this process in a totally coherent manner. They will learn and evolve to anticipate user requirements. According to the European Information Society Technologies (IST) Advisory Group, “The concept of Ambient Intelligence provides a wide-ranging vision on how the Information Society will develop. The emphasis is on greater user-friendliness, more efficient services support, user-empowerment, and support for human interactions. The Ambient Intelligence environment is capable of recognizing and responding to the presence of different individuals. And, most importantly, Ambient Intelligence works in a seamless, unobtrusive and often invisible way.” [1] Areas such as medical monitoring and telemedicine, automobiles,

sports, and entertainment are currently beginning to benefit from innovative applications that are building blocks for these AmI systems [2, 3].

Ambient intelligence can be represented as a technological convergence of ubiquitous computing, ubiquitous networking, and intelligent user interfaces, each of which has drivers in different research disciplines. Thus, the challenge of implementing the vision of AmI is so great that only a highly multidisciplinary approach will be successful. This aspect has been given a prominent profile throughout European research (the information society technologies, or IST, programme of the current Framework 6 is focused on creating AmI [4]), and a significant amount of descriptive work, in the form of scenarios, has been done to define the central barriers to progress.

From a technological perspective, concrete requirements for ambient intelligence are (1) very unobtrusive hardware, (2) a seamless mobile and fixed communications infrastructure, (3) dynamic, massively distributed networks of collaborative devices, and (4) intuitive, dependable human–computer interfaces that engender trust. Much of the focus of current AmI research is in software development and even concept design. However, significant challenges are present in hardware systems research also. For instance, the requirements for unobtrusive hardware contain a number of key research challenges [1], including:

1. self-generation of power and micropower-level performance;
2. smart surfaces and new I/O displays;
3. new active sensor and actuator devices and subsystems;
4. interfaces for user and environmental interaction;
5. smart materials that can change their characteristics;
6. nanoelectronics: development of nano–micro devices.

These challenges require significant progress in materials and process development, miniaturization, new transducer devices, and new integration techniques for scaling systems performance (both in terms of computational capacity and physical size). A major driver in this area will be the modularization of these capabilities, and their effective and appropriate introduction into new products. In terms of achieving the goal of AmI, this will be represented first by the development and introduction of foundation-level intelligent systems and products. These will seek to “seed” AmI systems. Further scaled deployment of hardware systems over time will then be followed by a hybridization process to link these together as coherent and adaptive user services. An example of such a process, given at technology platform and product level, is shown in Table 24.1.

24.2 The Role of Silicon

Hardware development in AmI will be built upon existing innovation routes, and based around silicon’s capability to deliver performance in accordance

Table 24.1. Technology challenges for implementing ambient intelligence [1]

Short-term technology issues (2005–2007)
Microsized software radio transceivers
Personal ID “key of keys”
Radio bandwidth of 20 Mbps and wireless protocol interoperability
Flexible user interfaces
Modular design of agents
Augmented objects (IP address and user interfaces)
Multifunctional personal identifier (biometry or chip implants)
Stand-alone microsized devices with data capture and ad-hoc wireless
Behavioural pattern recognition – upload of information into “knowledge”
Intercommunicating active (lights, etc.) and passive (labelling, etc.) devices
Intelligent agents that learn, infer, and negotiate
Transition from object to quasi-subject
Middle-term technology issues (2008–2010)
Personal area networks – wireless, bandwidth, scalability, miniaturization
Low-power sources and autonomous power supplies
Pico-radio technology – “short-hop” wireless
Smart materials embedded in both vehicles and wearable communicators
Advanced tagging systems and devices (RF ID tags the size of a grain of rice)
Long-term technology issues (2010–)
Tangible/tactile and sensorial interfaces
Miniaturization and nanotechnologies: very small cameras, sensors, and actuators
Miniaturization and nanotechnologies: combined with smart materials
End-of-life management of augmented objects

with targets set by Moore’s law. However, the drive to miniaturization and unobtrusiveness will require both the form and function of the silicon to adapt through increased hybridization with novel integration techniques, and with emerging platforms such as nanotechnology. In fact, the importance of this is such that hybridization of micro–nanosystems is seen as one of the three major breakpoints leading to implementation of AmI on a global scale. (The other two are the future emergence of open-standard software interfaces and the growth of fuzzy matching techniques for applications of artificial intelligence) [1]. The nature of the systems development will build upon current information systems technology research, which is silicon-centric. Novel materials, such as smart polymers, will be integrated into our surroundings to extend the scalability and versatility of the hardware in the face of a requirement for significant cost reductions. However, at the core of AmI is the effective deployment of systems intelligence. This is a capability that is driven by computationally intensive software algorithms, requiring significant scope for hardware functionality, and only silicon can foreseeably facilitate a timely deployment of effective AmI platforms.

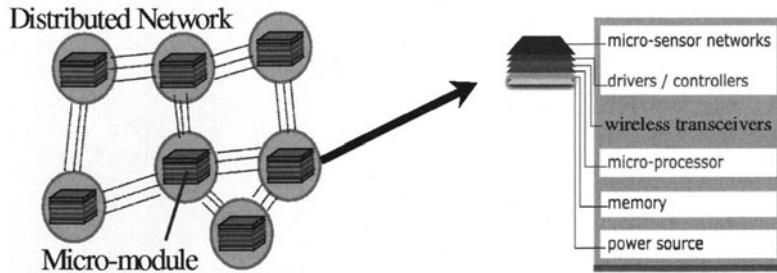


Fig. 24.1. Description of distributed systems, including a schematic of an individual module

24.2.1 Modular Computational Platforms

Recent developments in wireless and microsensor technologies have provided foundation platforms for considering the development of effective modular systems (see Fig. 24.1) [5]. They offer the prospect of flexibility in use, and network scalability. Currently, most sensor networks are strongly integrated into the assembly process of their target systems (e.g. the automobile, production line equipment, and aircraft). Thus, they carry a high infrastructural overhead. Emerging autonomous formats include wireless units designed to collect data and transmit to central (or distributed) hosts. Interesting examples include passive/active tags, inertial measurement units (IMUs), and 1 cm^2 wireless integrated microsensors being studied at UCLA [6, 7].

One of the most relevant research activities worldwide is the “Smart Dust” project at the University of Berkeley [8, 9]. In this project, the goal is to develop a system of wireless sensor modules where each unit is the size of a mote of dust. The work includes miniaturization, using die bonding, flip-chip, and wirebonding assembly sequences, integrated microsensors; and computation. It also includes wireless communication in a 1 mm cube. A key augmenting performance factor will be the incorporation of reconfigurable system-on-chip solutions driven by a requirement for dynamic field programmability. These platforms will facilitate the introduction of intelligent behaviour, through the placement of evolving genetic algorithms and novel neural networks on ASIC-format chips.

24.2.2 Microelectromechanical Systems

According to a forecast from Cahners In-Stat Group, microelectromechanical systems (MEMS) consumer electronic sales will surge from about \$3.9 billion in 2002 to over \$8 billion by 2007 [10]. Moving beyond applications in specialized markets, such as automotive (for airbag deployment) and medical (blood pressure sensors), MEMS are poised to infiltrate consumer electronics. Accelerometers, mirrors, relays, and other MEMS devices are moving

into production as they show a potential ability to improve the functionality and performance of products such as camcorders, personal digital assistants (PDAs) and digital versatile discs (DVDs). The most promising applications on the horizon include accelerometers in hard disk drives, microrelays in cellular phones, and digital micromirror (DMD) devices in TVs, home theatre systems, and projection displays. In addition, future applications will likely include motion/vibration sensing, fingerprint recognition, mass storage, and signal switching. There is a great deal of interest in manufacturing processes that allow the monolithic integration of MEMS with driving, controlling, and signal-processing electronics. Integration promises to improve the performance of MEMS devices, as well as the cost of manufacturing these devices, by encompassing the sensor device and its electronic subsystem in the same manufacturing and packaging process.

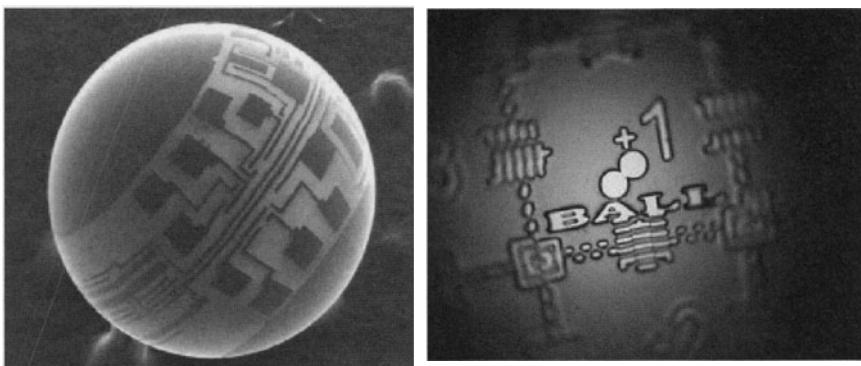


Fig. 24.2. A spherical IC manufactured by Ballsemi, showing a full 1 mm sphere, and a magnified image of the etched metallization pattern

24.2.3 New Silicon Form Factors

More and more active silicon devices are being developed using non-planar fabrication techniques. A notable example is the research work performed by Ballsemi in developing 1 mm spherical integrated circuits (see Fig. 24.2) [11]. The advantage provided by using the spherical shape is an increased available surface area, and the potential for embedding microsensors in the core of the sphere, thereby optimizing the performance of the silicon volume. Current development work is focusing upon the fabrication and development of autonomous sensor spheres for medical monitoring applications.

The drive to lower-profile electronics systems, as represented by the Smart Card market, has lead to research on thin silicon devices, yielding promising behavioural characteristics, and methods for very high-density integration. The low-stress dicing by thinning (DbyT) process developed by Fraunhofer-IZM has successfully demonstrated high levels of mechanical flexibility in

silicon devices thinned to the order of 20 µm [12]. Thinning techniques such as back-surface grinding to 50 µm are becoming common, and are viable in certain cases for building ultrathin 3-D stacked silicon systems. Companies such as Tru-Si have accelerated this process by developing the atmospheric downstream plasma (ADP) process, which can be used to thin silicon and to enable their Thru-SiliconTM process, which creates through-silicon vertical interconnections between the front and back surfaces of a wafer in a manner similar to that of through-holes in printed circuit boards [13].

These methods will be incorporated into the developmental methodology for creating AmI, but are essentially driven by current IT trends. Techniques that are specifically focused upon AmI, while seeking to take advantage of the mechanical behaviour of silicon, include the silicon “Fibre Computing” technique [26], which is developing methods to fabricate flexible ICs that can ultimately be embedded into fibrous materials, including textiles for wearable computing.

24.3 Ambient Intelligence: Developmental Methodology

Major research efforts are currently targeting the “disappearance” of the computer into the fabric of our environment. In the future, the spaces in which we live will be populated by many thousands of objects (often described as “artefacts”) with the ability to sense and actuate in their environment, to perform localized computation, and to communicate, even collaborate, with each other. Augmented everyday artefacts are playing a large role in research towards intelligent systems and ubiquitous computing. There are two prime drivers: the smaller these smart artefacts are, the more effective they will be in creating coherent AmI systems, and the greater the number of objects within these systems and networks, the more valuable these networks become.

The main properties required to maximize the capabilities of such networks are that they should have scalability (i.e. the number of nodes can be increasingly large) and granularity (i.e. high resolution). They should also be reconfigurable (i.e. capable of self-organization), modular (i.e. they should allow ad-hoc interconnection and autonomous behaviour), and mobile (i.e. they should have physical and digital portability) [14]. In order to realize truly intelligent systems, there are a number of basic requirements: the system must be able to sense its surroundings and act upon what it perceives. As the number of sensors increases, so too does the complexity of the computation and communication by the systems that is required. When the number of sensors in a local area is extremely large, there will be scaling effects that are currently unforeseeable, and potentially beneficial emergent behavioural effects will be engineered from this. Implementation will be realized through concurrent hardware–software engineering; innovation in software should be matched by invention in hardware.

The key elements for achieving the hardware-oriented objective of miniaturization are: (1) novel computational solutions and systems that address power-aware computation, power-aware communication, and the scalability of the system itself; (2) novel advanced interconnect and packaging technologies to achieve the required miniaturization in a manufacturable, cost-effective way; and (3) new silicon formats and devices. It is important that novel hardware technology platforms are used for object and system development, incorporating 3-D stacking, multichip and microsensor integration, thin and flexible IC substrates, active polymeric materials, smart materials, and ultimately micro–nano hybrid systems. To do this, new form factors for hardware need to be investigated, allowing a synergistic effect to emerge at an application level that optimizes overall performance. In this light, the key initial considerations are interconnection and modularity of the hardware. For AmI, this is primarily necessary in order to embed electronics into everyday objects. However, explicit deployment of miniaturized autonomous modules into, for example, harsh environments could also have a very high market potential.

24.4 Novel Computational Solutions and Systems

Notable development work for distributed collaborative systems is taking place in the research area of robotics [15], and in the “Disappearing Computer” initiative [16]. Artificial intelligence methods, employing neural networks and genetic algorithms are typical tools for robotics research; in this case, as size limits are not a major concern, issues of power-aware computation and communication are of lower priority. Scaling such systems has proved highly complex. Similar techniques are being applied to the development of solutions for distributed artefacts, which are everyday objects with embedded electronics. The nature of the individual artefacts is based upon mobile computational units, such as the mobile phone, and thus intelligent use of power and bandwidth are major issues. Again, the process of scaling is a major issue, in particular where dynamic, mobile services and functions are required of the distributed system of artefacts. The use of embedded component software, integrated into objects to give them both a physical and a digital presence, is one method under development to address this issue. In fact, creating innovative (top-to-bottom) open-standard architectures is perhaps the central focus of all of the current and ongoing research on AmI and ubiquitous systems.

24.4.1 The Disappearing Computer

In Europe, a focal point of this research is the “Disappearing Computer” initiative [16]. The goal of this programme is to explore how everyday life can be supported and enhanced through the use of collections of interacting artefacts. The initiative has three interlinked objectives:

1. Develop new methods for the embedding of computation into everyday objects.
2. Perform research on how new functionality can emerge from interacting artefacts.
3. Ensure that people's experience is coherent and engaging in space and time.

The programme consists of 17 research projects, each addressing aspects of the above objectives. Among them is the "Extrovert Gadgets" project [17], which aims to provide a conceptual and technological framework that will engage and assist ordinary people in composing systems of computationally enabled everyday objects.

24.4.2 Extrovert Gadgets

The objective of this research project is to develop a method of diffusing information technology into everyday objects and settings, and assisting ordinary people in composing, (re)configuring, or using groups of these objects. To achieve this, the project has developed and validated a software framework, known as a *gadget-ware architectural style* (GAS). This constitutes a generic framework shared by both gadget designers and users for consistently describing, using, and reasoning about families of augmented objects (or "eGadgets"). Generally speaking, eGadgets are everyday tangible (physical) objects enhanced with sensing, acting, processing and communication abilities. This may entail "intelligent" behaviour, which can be manifested at various levels. In effect, eGadgets are "GAS-aware" artefacts, which are used as the building blocks of "Gadget-worlds". Gadget-worlds are dynamic functional configurations of egadgets exhibiting collective behaviours (the equivalent of software applications consisting of interacting objects).

The current system is fully effective in demonstrating that the concept of dynamic reconfigurability of component systems can be applied to everyday objects [18]. Groups of eGadgets can be "plugged" together to give numerous types of functional gadget-worlds (see Fig. 24.3). In addition, the system shows an encouraging level of longevity in its performance. From the perspective of transducer function, the system operates acceptably within the confines of the dedicated scenario. However, there are relationships between the context of the use of the gadget-worlds and the nature of their transducer networks that requires further study [19].

24.5 Novel Advanced Integration Technologies

An integral part of achieving unobtrusive systems is through miniaturization. In general, the smaller the gadget the more difficult it is to effectively embed

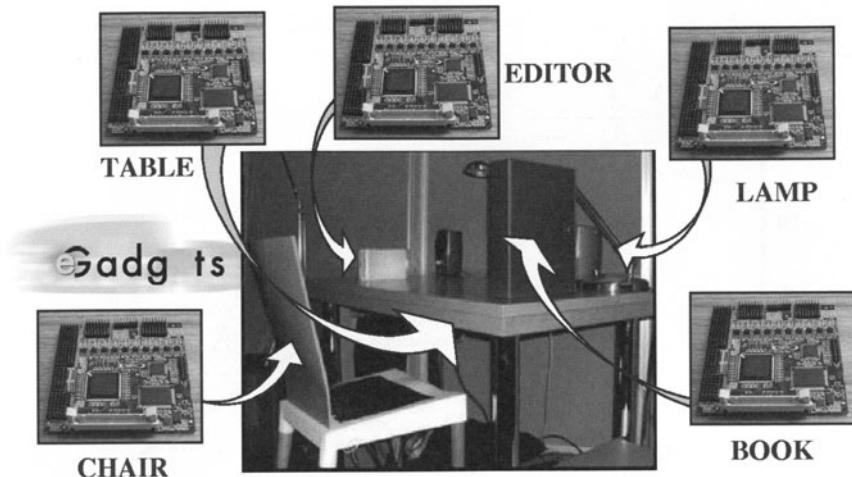


Fig. 24.3. The construction of a study gadget-world in which sensor systems are embedded in numerous objects, which can be digitally “plugged” together to perform user-defined functions such as controlling and automating lighting conditions during study periods

sufficiently dense transducer granularity levels for good performance. However, typically, this is not due to the size of the transducer. It is due rather, to the protective encapsulation materials and the interconnects (including on-chip metallization) in use to provide traditional solutions for handling, connectivity, and deployment. Device and component miniaturization is in part solving this problem; however, there is a trade-off between device size and system scale, as the complexity of the conditioning and aggregation layers tends to increase significantly. Current high-density interconnect solutions are not easily designed to supply connectivity over (in relative terms) large surface areas, and, more importantly, are extremely difficult to deploy effectively. In the context of creating AmI, these issues become significantly complex, and require convergence of strategic research to be implemented. One hardware-focused example is NMRC’s Intelligent Seed programme [14].

24.5.1 Intelligent Seed Programme

The Intelligent Seed programme is about creating appropriate functional architectural templates for enabling AmI, and providing appropriate levels of miniaturization that solve future production and usability issues. The primary hardware research target that connects these goals is a requirement to develop very highly miniaturized wireless microsensor networks of the order of 1 mm^3 in size (see Fig. 24.4). This programme is guided by a roadmap that incorporates the development of 25 mm (see Fig. 24.5), 10 mm, 5 mm, and 1 mm cubic modules, each undergoing development and evaluation through

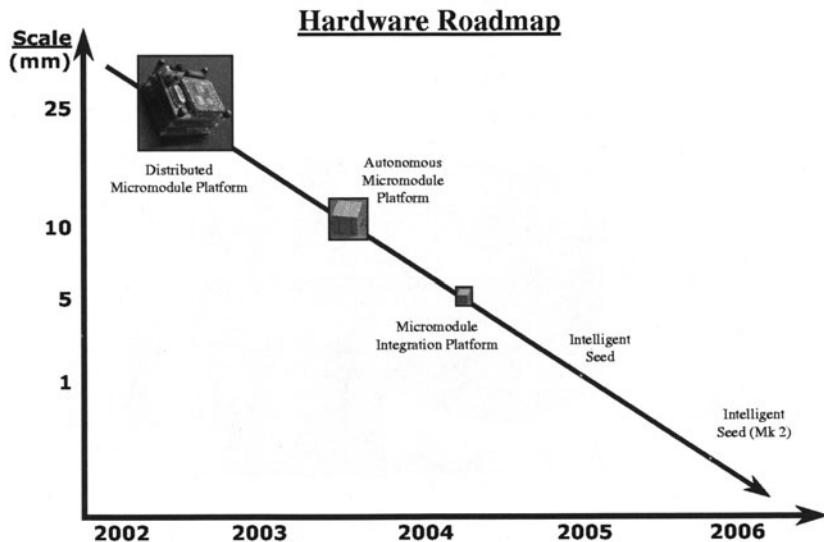


Fig. 24.4. Hardware roadmap detailing miniaturization targets for the Intelligent Seed research programme [14]

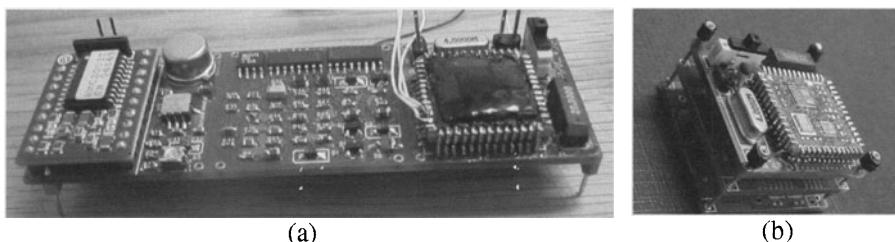


Fig. 24.5. Autonomous sensor network node with modular 25 mm cubic design format (a) before, and (b) after module segmentation and 3-D stacking [20, 21]

selected experimental AmI scenarios. The research is directed towards creating globally scalable, high-granularity technology platforms that can be deployed easily and effectively into many types of user environments or embedded into the everyday objects that are used in these environments.

The larger 25 mm and 10 mm cubic formats provide the scope for the rapid prototype assembly of autonomous transducer systems, serving a critically important purpose in merging hardware, software, and user-design research. Each of these module forms can be viewed as miniaturized, reconfigurable laboratories, and thus as focal points for use in investigating issues such as sensor architecture (e.g. for effectively implementing extrovert gadgets), systems deployment, ad-hoc networks, the integration of intelligence, and power management and generation. In terms purely of hardware development,

these are key factors needing definition, at least as part of a dynamic, iterative process in order to create appropriate targets for functional miniaturization.

The highly miniaturized form factors of the 5 mm and 1 mm Intelligent Seeds utilize numerous high-density integration techniques that are the subject of current state-of-the-art research worldwide. In particular, these include multichip modules (MCMs), 3-D integration, and development of the system-on-chip (SoC) and system-in-a-package (SiP) technologies [22–24]. Material systems and processes are also being adapted to encompass the changing requirements for very highly integrated autonomous systems. These include the development of very thin flexible multilayer substrates (of the order of 2–3 μm per layer), the integration of thin silicon ICs (each at 50 μm or less), and the development of novel substrates [25] – in particular silicon fibre computing [26] – and new approaches to solve handling and assembly issues for Intelligent Seed modules. The following sections describe these integration and substrate development approaches in more detail.

3-D Packaging Techniques

As demand increases for low-cost electronic products that offer smaller size, higher performance, and increased functionality, packaging technology has been required to improve significantly to keep up with the market expectations. One of the relevant developments in packaging to meet this growing demand was chip-scale packaging [22]. These miniaturized packages combined the benefits of flip chip (i.e. shorter interconnections, and area array connections) with the considerable benefits of the package itself (i.e. physical protection and access to volume-scale surface mount (SMT) assembly). These methods have evolved into 3-D packaging (i.e. vertical stacking), which has now emerged as an innovative way to meet the market requirements for new generations of electronic products [27]. By stacking die, it is potentially possible to improve heat dissipation and system integration while also reducing the length of the wire connections between the die, thus reducing noise and increasing speed. One of the biggest advantages of 3-D packaging is that it increases silicon area efficiency – the ratio of the total substrate area in the stack to the footprint area – making values of greater than 100% possible [28].

In fact, the idea of stacking chips is not very new. Sharp Corporation led the way in 1998, when it introduced the first stacked chip-scale package (S-CSP) of bare-die flash and SRAM for cell phones [29]. Today, Fujitsu, Hitachi, Mitsubishi, NEC, ASE, Toshiba, Dense-Pac Microsystems, and Amkor Technology are among the companies producing different kinds of S-CSPs for portable devices [27]. In a typical package of this kind, two or three memory chips are piled on top of each other, separated by a thin layer of die attach material and connected by wires to die-bond pads on the package substrate. There are a number of companies producing truly innovative 3-D assembly techniques. Tessera Technologies has a method for attaching memory chips and ASICs to flexible substrates and folding them over to create

low-profile stacks [30]. Valtronic SA of Switzerland already folds logic and memory components into a single package for hearing aids and other low-volume, high-value applications [31]. On the horizon is a system in a cube, based on epoxy-moulded layers of different chips. Developed primarily for military and space applications by companies such as Irvine Sensors Corporation [32] and 3D-Plus [33], these superstacks are now found in microcameras for satellites.

There are three basic methods of 3-D stacking, which can be classified as wafer-level, chip-level, and package-level stacking.

Wafer-level packaging [34], which is still in its infancy, is most promising and involves stacking wafers on top of each other. Interconnection is made through via-holes opened in the wafers. A schematic is shown in Fig. 24.6.

Chip-level stacking [35] involves stacking chips on top of each other, and interconnections are obtained by either wire-bonding or flip chip techniques. This is one of the most documented and researched stacking techniques, and memory chips and hearing aids are manufactured commercially in this way. An example is shown in Fig. 24.7.

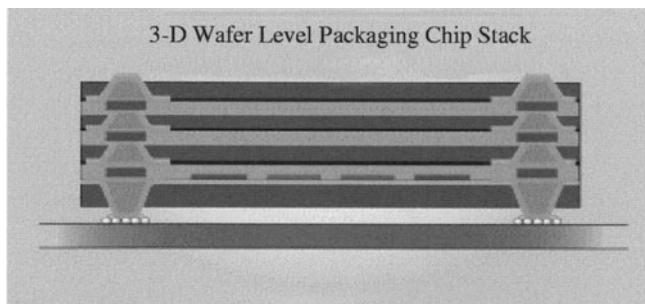


Fig. 24.6. Schematic illustration of wafer-level packaging, showing different wafers stacked on top of each other, with interconnections formed through the via-holes in the wafers [34]

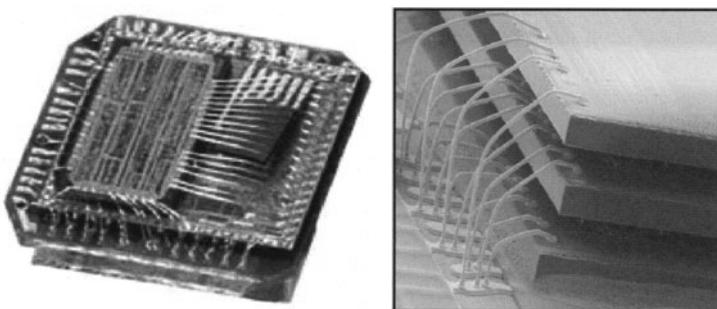


Fig. 24.7. Chip-level stacking, showing three chips stacked and interconnections made through the wire bonding [35]

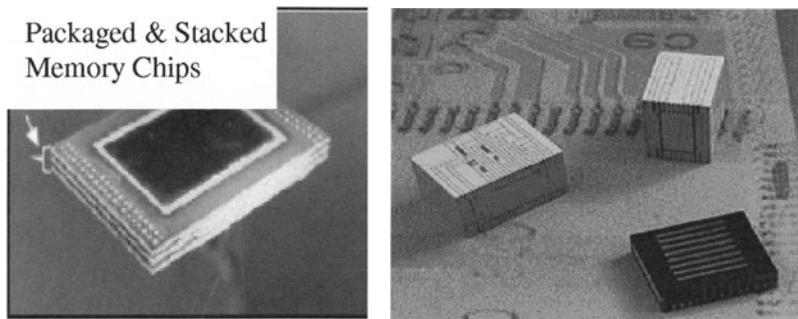


Fig. 24.8. Packaged IC stacked using rigid and flexible printed circuit boards, and the three-dimensional package made by 3D-Plus [36]

In package-level stacking [33], plastic packages are stacked to increase the silicon efficiency. The advantage of this technique is that the devices have already been tested, showing increased packaging reliability. Package-stacking technology is currently available in a stacked thin small-outline package (TSOP). An example of a plastic package is shown in Fig. 24.8, in which very thin packages, called paper-thin packages, are stacked on top of each other.

3-D Packaging for the Intelligent Seed

Moving from planar and 2-D fabrication processes to 3-D methods is a central part of creating the level of miniaturization necessary to implement autonomous wireless micromodules. The nature of the stacking approach is inherently modular, and, as a result, it is possible to develop system-on-a-chip and system-in-a-package versions that are stand-alone nodes within an autonomous network. It is possible to integrate the entire functionality of a wireless autonomous sensor node into a stacked package while continuing to maintain a level of functional flexibility suitable for a broad range of applications.

The hardware miniaturization process for the Intelligent Seed applies numerous enabling technologies for high-density integration where the packaging material is largely removed, and the targeted form factor of the modular units is cubic. The enablers include flip chip techniques, assembly of functioning thin silicon circuits, fabrication of stacked IC modules (including microsensor and smart materials), and 3-D assembly techniques. There are significant difficulties in realizing a process for fabrication that provides for high-density interconnects between ICs within the module. However, a number of useful seed platforms exist. These will include silicon-thinning processes combined with through-silicon metallization and ultrathin multilayer flex technologies that can be wrapped around silicon devices. These seed technologies enable innovative concepts to be implemented to achieve the required interconnect form and density.

An example of such a concept is to combine through-silicon via interconnection of multiple layers of thin ICs with assembly on, and folding of, thin, flexible polyimide substrates. As the chip thickness decreases below 50 μm , the volume of the interconnect becomes a major proportion of the overall thickness of the 3-D package. If left alone, this has implications for the functional efficiency of the autonomous module. Thus, there is a need to reduce the interconnection material volume in line with the silicon volume. Creating via interconnect through silicon layers has the advantage of significantly reducing packaging materials; however, there is a trade-off between active silicon and 3-D interconnection through the silicon that may impact negatively on the computational power. The use of thin flex materials provides the prospect of significant reduction of the packaging and integration materials, while maximizing the effectiveness of the silicon layers.

A typical process of formation of a thin flex substrate starts with spinning polyimide onto a wafer and sputter-coating the polyimide with metal, followed by pattern generation and a second layer of polyimide and metal. One of the main challenges in the process is removal of this substrate from the wafer without subjecting the substrate to undue stress. Once the thin flex is obtained, thinned silicon chips are flip-chipped or connected to the substrate via various techniques. The flex is then folded several times to obtain a thin-profile 3-D package.

While the feasibility of these approaches has been established, significant research is required in order to determine appropriate techniques for functionally integrated autonomous wireless modules, and still more work is required to achieve scalable systems-level deployment. From a processing point of view, the handling of thin silicon (particularly wafers below 100 μm) and thin flex is a major challenge. Low-stress thinning and dicing techniques for ICs exist, but more research needs to be undertaken to understand the requirements for performance optimization. Little work has yet been done on the performance implications for microsensors in this regard. Investigation has shown that the device characteristics of thin silicon ICs are reproducibly affected by the influence of bending stress.

The development of thin flex interconnects is promising. However, the handling of these flex materials is at least as challenging a problem as that of thin silicon wafers, primarily because of their tendency to stretch or wrinkle given small variations in handling stresses. Additionally, there will be design issues, as well as materials and process development requirements, that arise from the need to offset the natural rigidizing effect which occurs as the number of flex layers is increased. The effects of thin flex on the electrical, thermal, and mechanical characteristics of the modules will need to be better understood to assess the technique for various AmI applications.

24.6 New Silicon Forms for Sensors and Actuators

There are numerous programmes of research undertaking the development of novel distributed sensor systems [37]. These programmes are primarily software-focused, concerned with information management and the reliability of scalable systems, or concerned with effective power management. The majority of this research has been performed using modelling techniques, or through laboratory-level implementation of macrohardware systems. The transition to genuine scalability will require significant innovation in systems integration for the physical networks themselves, particularly in developing effective processes and materials for fabricating and integrating the sensor devices.

One area requiring significant applications-directed research is that of wearable computing. The concept of wearable computing opens up entirely new possibilities in areas such as medical monitoring and telemedicine, sports and athletics, entertainment, and expressive musical/dance performance [38]. It offers the potential for a very high degree of applicable functional processing power for the user, provided in a particularly convenient manner through placement of non-invasive sensors around the body. A key barrier to this is unobtrusive deployment of effective sensors.

A project within the Disappearing Computer initiative, known as “Fibre Computing”, is developing techniques to integrate computing ability directly into fibres [39]. The project goal is to fabricate electronic components and circuits into fibres, which can ultimately create objects (ranging from clothing to carpets) that can interact with each other or with their surroundings. Clothes, furniture, and many other products can then be woven from these flexible, functional fibres.

24.6.1 Fibre Computing Technology

Much work has been done on the examination of the strength of silicon microstructures and it has been established that silicon structures become extremely flexible when sufficiently thin [40, 41]. One application for this is in the form of a functional silicon fibre to be used in “Fibre Computing” [26], which has the potential to change the way advanced circuits and systems are designed and fabricated in the future and is potentially the next evolutionary step in wearable computers. The research aims to make large, flexible integrated systems for wearable applications by building functional fibres with single-crystal silicon transistors at its core.

A novel technology that enables the fabrication of functional flexible silicon fibres has been developed within the Fibre Computing research project [42]. The concept involves building a circuit in silicon-on-insulator material (SOI), laying it out in such a way that its topography is linear and constrained in the direction of the wafer diameter, and releasing the circuit by undercutting the silicon dioxide layer by means of a chemical wet-etch

process. This leaves the fibre completely free to move and to be extracted onto a support structure. A functional device based on this experimental approach has been successfully fabricated in the form of a functioning PN junction. Subsequent active device circuits have been designed and are currently being fabricated. This technique has the potential to provide a planar technology that can manufacture extremely powerful circuits and systems in long, narrow fibres, which can be woven into fabrics.

Mechanical Design of Silicon Fibres

To investigate the influence of the mechanical strength of the fibre shape, 64 different structures were laid out on separate die on a test wafer. The three basic test structures designed to evaluate the mechanical properties of the fibre were the S, C, and chair shapes.

The parameters for the silicon fibres were length (L) 200–2000 μm , width (W) 75–150 μm , and radius of curvature (R) 100–150 μm (see Fig. 24.9). The final fibre design was an array of these shapes linked together and could be between 4 mm and 42 mm long. A practical example of the final test structure

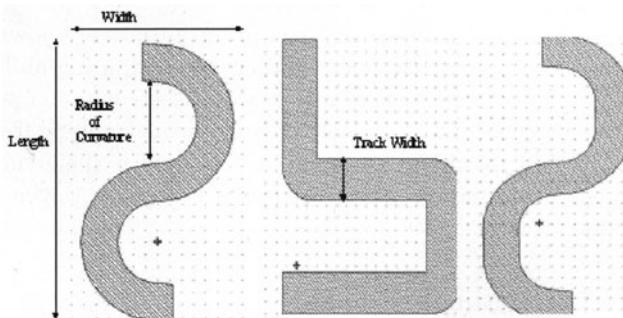


Fig. 24.9. Test structures: C, chair, and S, from left to right

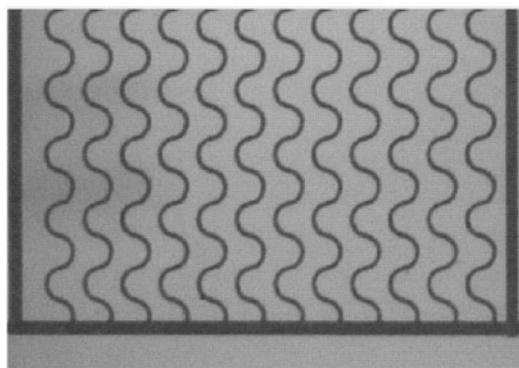


Fig. 24.10. Test structure with 100 μm frame around the perimeter of the fibres

is shown in Fig. 24.10. Between each of the fibres is a spacing of 30–35 μm , and the track width for each is between 30 and 35 μm . Around the perimeter of the fibres there is a frame 100 μm wide to support the structures during removal from the handle wafer. The frame was removed when testing was performed.

Fabrication of the Fibres

To examine the feasibility of producing a functional silicon fibre, SOI-type wafers were fabricated [43]. Beginning with a $\langle 100 \rangle$ silicon wafer 525 μm thick, a 2 μm polycrystalline silicon layer was deposited on top of a 1 μm thermally grown SiO_2 layer. The polycrystalline silicon was patterned in the desired shape of the fibre, and then subjected to diffusion doping with phosphorus (implantation of $5 \cdot 10^{15}$ atoms cm^{-2} at an ion energy of 100 keV) and boron (implantation of $5 \cdot 10^{13}$ atoms cm^{-2} at an ion energy of 45 keV) to create the necessary metallurgical junction. An illustration of the fabrication process can be seen in Fig. 24.11. The wafer was diced and the fibres were released from the individual die by undercutting the sacrificial silicon dioxide layer using a 5:1 buffered oxide chemical wet etch. Fibres were manually removed from the etchant and stored in a gel pack container.

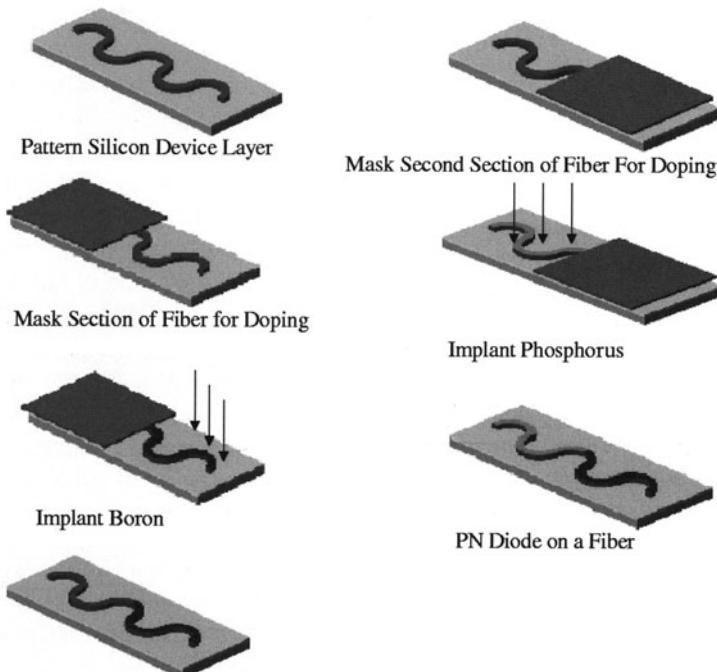


Fig. 24.11. Fabrication procedure for the fibre test structures

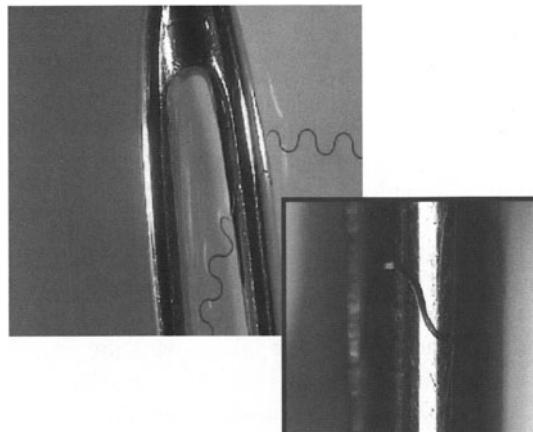


Fig. 24.12. Free-standing fibre after release process being threaded through the eye of a needle

Mechanical and Electrical Evaluation

The C-shape test structure proved to be the most structurally reliable of the three designs. Although no conventional stress measurements were performed on the free-standing fibre, the flexibility and mechanical robustness were observed during the release process and handling of the fibres. However, the frame designed to give the fibres mechanical stability did not lift off from the carrier substrate and the fibers had to be manually cut from the frame. It is unknown whether this was due to an etch problem or stiction issues [44]. This produced a collection of individual, randomly dispersed fibres floating in the etchant, which had to be manually retrieved. Retrieving the fibres manually caused them to rotate, and identifying the doped side of the fibre for subsequent probe testing was difficult. However, it was thought that there would be a Gaussian doping profile for such a thin film, giving a uniform distribution on both sides. A fibre after release is shown in Fig. 24.12.

Electrical testing of the devices proved to be extremely difficult owing to the flexible nature of the unconfined fibre. Also, the dimensions were similar to that of the probe tip, making the positioning of the probe difficult. However, the first electrical device on a semiconducting fibre was produced, and its diode characteristics before and after release are illustrated in Fig. 24.13.

The lack of continuity between the curves for before and after the release process could be related to the amount of time the fibre remained immersed in the etchant and the random bending of the crystalline microstructure on removal from the etchant. After the concept of flexible active devices was established, a more complex demonstration of functionality was investigated. Subsequent active device circuits have been designed and are under fabri-

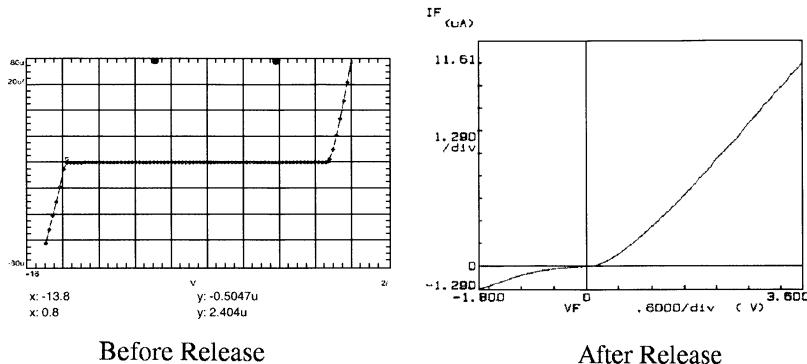


Fig. 24.13. Characteristics of fibre diode before and after the release process

cation. Results from a contact chain based on the SOI technology will be examined in the following sections.

Design of Active Device Circuits

The ring oscillator is a standard circuit for delay measurement. The layout consists of an odd number of inverters (see Fig. 24.14) connected in a circular chain. For the purpose of this experiment, a 679-stage ring oscillator was designed.

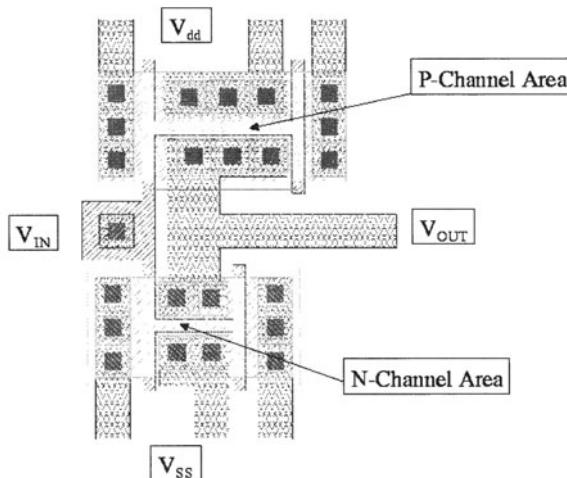


Fig. 24.14. Design layout of an SOI inverter

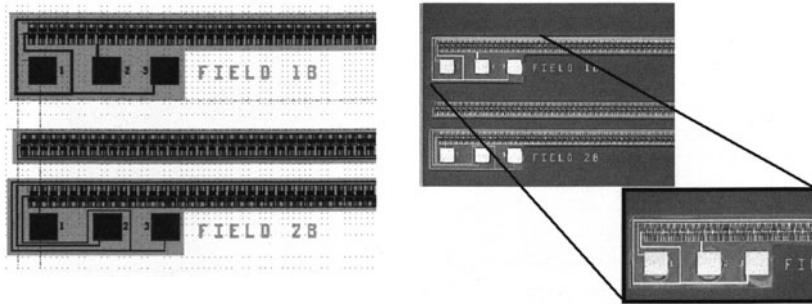


Fig. 24.15. Design layout of a ring oscillator, and final circuit after $1.5\text{ }\mu\text{m}$ CMOS processing

The mechanical flexibility required to integrate the circuit into a wearable artefact is a key issue. This has been resolved initially by using a polyimide flex material instead of BPSG oxide as the interlayer dielectric in the circuit [45].

Secondly, when polyimide is used as a passivation layer covering the circuit, it will act as a support structure holding all the individual silicon islands along the circuit together after release from the handle wafer. There are three different types of ring oscillator that have been designed, all with different numbers of chains of inverters involved, with each chain consisting of 679 stages. An example of the ring oscillator designs is illustrated in Fig. 24.15.

Fabrication of Active Device Circuits

The ring oscillator was fabricated on single-crystal silicon and polycrystalline-silicon SOI-type wafers. The handle wafer was $525\text{ }\mu\text{m}$ thick with a 4000 \AA buried oxide layer and a 3400 \AA thick silicon device layer. A brief outline of the fabrication process can be seen in Fig. 24.16. Firstly, the silicon islands are defined by means of a plasma etch, and the gate oxidation and well implants follow. The N and P well implants are split into two, with the N well having a deep phosphorus implant ($3 \cdot 10^{12}\text{ P}^+ \text{ cm}^{-2}$ at 190 keV) and a top boron implant ($2 \cdot 10^{12}\text{ B}^+ \text{ cm}^{-2}$ at 20 keV). The P well implant is split between a deep boron implant ($2 \cdot 10^{11}\text{ B}^+ \text{ cm}^{-2}$ at 70 keV) and a top boron implant ($1.1 \cdot 10^{12}\text{ B}^+ \text{ cm}^{-2}$ at 20 keV). A 3500 \AA layer of polysilicon is deposited and patterned to create the gate. This is followed by source/drain implants of phosphorus ($5 \cdot 10^{14}\text{ P}^+ \text{ cm}^{-2}$ at 60 keV) and boron ($2 \cdot 10^{11}\text{ B}^+ \text{ cm}^{-2}$ at 70 keV). The contact stage is a $3\text{ }\mu\text{m}$ patterned layer of polyamide, used to increase the flexibility of the circuit after release from the handle wafer. A 6000 \AA layer of Al/Si1% metal is deposited and patterned to create the interconnect between silicon islands. Finally, a polyamide passivation layer is deposited and patterned over the circuit to increase circuit flexibility and

overall mechanical robustness. The final circuit before polyamide passivation processing can be seen in Fig. 24.16.

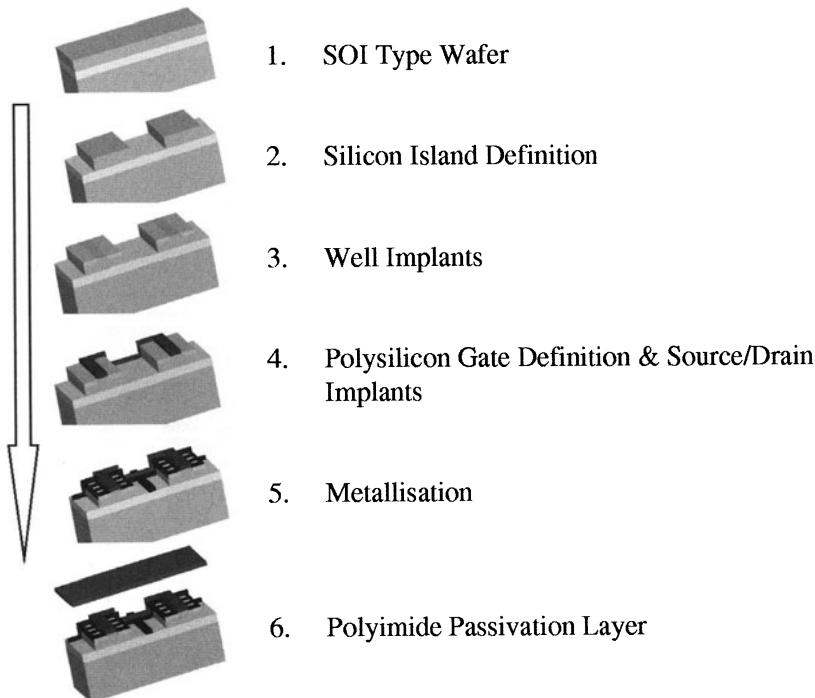


Fig. 24.16. Process flow for active device circuit

Evaluation of Silicon Fibre Circuit

To test the electrical performance, a short-loop process, which consisted of an array of metal-to-polysilicon contact chains based on SOI technology, was developed. This process included the novel polyimide inter-dielectric layer used in the fabrication of the device circuits. Figure 24.17 illustrates the results obtained for a 1359-stage metal-to-polycrystalline-silicon-island contact chain. From this experiment, similar results would be expected for the ring oscillator circuit.

Future Developmental Targets for Silicon Fibres

Future work will involve the release of the active fibres from the handle wafer and the subsequent testing of the free-standing active device circuits. The individual fibres will be released from the handle wafer by a combination of

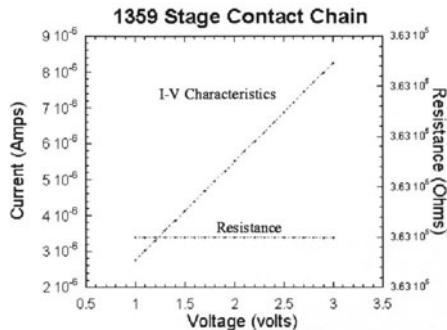


Fig. 24.17. $I-V$ characteristics and resistance measurements of metal-to-polysilicon-island contact chain

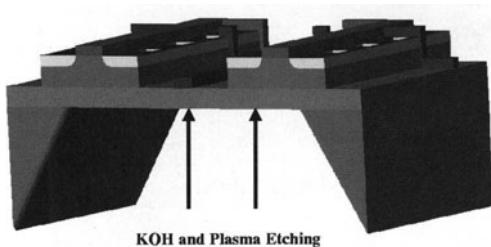


Fig. 24.18. Schematic illustrating one possible circuit release process

KOH wet etching and dry plasma etching of the bulk silicon through the back of the wafer, using the thin oxide layer as an etch stop (see Fig. 24.18). The fibres will be cut from the oxide by means of a focused ion beam.

24.7 Conclusions

Ambient systems are systems that will use electronics with a learning capability to proactively and seamlessly support and enhance our everyday lives. In this regard, AmI is an extremely user-focused research topic. From a technical point of view, creating AmI environments means integrating and networking numerous distributed devices in the physical world: in workspaces, hospitals, homes, vehicles, and the environment. Such progress will open up entirely new possibilities for future applications and resultant markets. Areas such as medical monitoring and telemedicine, sports, and entertainment are currently beginning to benefit from research that is attempting to develop building blocks for these systems. However, the process is highly complex and involves numerous challenges that may be solved only through strongly multidisciplinary methods. As a result, experts in this research area take an extremely highly iterative approach to investigative development. Many

innovative high-density interconnect (HDI) technologies are viable tools for implementing these systems through the fundamental activity of embedding electronics into everyday artefacts. Research into the issues of design, materials, and process development for multichip packaging (MCP), direct-chip-attach, 3-D circuits, thin silicon, and flexible multilayer substrates will in all of these cases facilitate progress. However, success will require a user-centric approach that challenges many aspects of the existing materials and manufacturing formats.

Much of the hardware development research performed in the area of ambient systems uses conventional high-volume assembly techniques. This will continue to be the case; many of the fundamental issues can be successfully investigated using these formats. However, there is a major transformation required in moving from effective prototyping techniques at laboratory level to providing genuinely scalable and deployable products, and development of the correct hardware formats is at the centre of this challenge. Traditional technology formats will not be effective on their own, and even state-of-the-art processes such as multichip modules, 3-D devices, and thin flexible circuits will struggle to meet the requirements. For scalability, more evolved solutions using innovative microelectronics (for example, silicon fibres, active polymers, and possibly 3-D silicon) and nanoelectronics will be necessary. A good example of what is required in this regard is represented by the implementation of the Intelligent Seed programme at NMRC, an initiative designed to effectively hybridize integration techniques in order to facilitate effective research into AmI platforms, and to provide solutions to genuine barriers to hardware development, through the creation of highly inventive fabrication techniques. One such technique is the flexible silicon fibre, a technology platform developed for wearable computing applications, which shows a high degree of potential for integrating functional sensors into everyday objects.

The challenge required to implement AmI is significant, and while the current focus on software innovation is necessary, the level of research required in order to implement effective hardware is very high. Silicon-based systems will be at the heart of this research, and there will be a requirement, in particular, for new forms of silicon substrates to be developed to volume-scale production in order that AmI may ultimately be realized.

References

1. IST Advisory Group (ISTAG), Scenarios for Ambient Intelligence in 2010, <http://www.cordis.lu/ist/istag-reports.htm>
2. C. O'Mathuna, K. Delaney, J. Alderman: MEMS packaging for the intelligent environment. Invited presentation at the IMAPS Advanced Technology Workshop, San Jose, November 9–12, 2001
3. J. Paradiso, K. Hsiao, A. Benbasat, Z. Teegarden: Design and implementation of expressive footwear. *IBM Systems Journal*, Vol. 39, Nos. 3 & 4, October 2000, pp. 511–529

4. The Framework 6 Programme – European Research Area, www.cordis.lu
5. K. Delaney, J. Barton: The challenge of globally embedding functional electronics into everyday artefacts. Conference on Smart Packaging, November 14, 2001, Institute of Materials, London, UK
6. D. Estrin R. Govindan, J. Heidemann, S. Kumar: Next century challenges: scalable coordination in sensor networks. ACM Conference Proceedings on Mobile Computing (Mobicom) (1999)
7. G. Pottie, W. Kaiser: Wireless integrated network sensors. Communications of the ACM, Vol. 43, May 2000, pp. 51–58.
8. B. Warneke, M. Last, B. Leibowitz, K. Pister: Smart dust: communicating with a cubic-millimeter computer. Computer, Vol. 34, No. 1, January, 2001, pp. 43–51
9. J.M. Kahn, R. Katz, K. Pister: Mobile networking for smart dust. ACM/IEEE International Conference on Mobile Computing and Networking (Mobicom 99), Seattle, WA, August 17–19, 1999
10. Cahners In-Stat Group, Got MEMS? 2003 industry overview and forecast, Research Report, 2003
11. Ball Semiconductor, Inc., www.ballsemi.com
12. K. Bock, M. Feil, C. Landesberger: Thin chips for flexible and 3-D systems. In: *Foldable Flex and Thinned Silicon Multichip Packaging Technology: IMAPS Emerging Technology in Advanced Packaging Series*, ed. by J. Balde (2003), chap. 5
13. Tru-Si Technologies Inc., www.trusi.com
14. K. Delaney, S. Bellis, S.C. O'Mathuna: Applying packaging innovation to ambient intelligence systems platforms. Proceedings of the IMAPS Nordic Conference, 21–24 September 2003, Helsinki University of Technology, Espoo, Finland
15. H. Hagras, M. Colley, V. Callaghan, M. Carr-West: Online learning and adaptation of autonomous mobile robots for sustainable agriculture. International Journal of Autonomous Robots, Vol. 13, July 2002, pp. 37–52
16. The Disappearing Computer initiative, <http://www.disappearing-computer.net/>
17. Extrovert Gadgets (e-Gadgets) website, <http://www.extrovert-gadgets.net>
18. A. Kameas, S. Bellis, I. Mavrommatti, K. Delaney, A. Pounds-Cornish, M. Colley: An architecture that treats everyday objects as communicating tangible components. IEEE International Conference on Pervasive Computing and Communications (PerCom2003), Dallas-Fort Worth, Texas, March 2003
19. K. Delaney, S. Bellis, C. O'Mathuna, A. Kameas, I. Mavrommatti, M. Colley, A. Pounds-Cornish: Unobtrusive transducer augmentation of everyday objects for systems with dynamic interactivity. Sensors & Their Applications XII, 2–4 September 2003, University of Limerick, Ireland
20. J. Barton, K. Delaney, S.C. O'Mathuna, J.A. Paradiso: Miniaturised modular wireless sensor networks. IEEE Conference on Ubiquitous Computing (Ubicom) 2002, September 29–October 1, 2002, Göteborg, Sweden
21. J. Barton, K. Delaney, N. O'Mahony, S.C. O'Mathúna: Functional integration for embedded intelligence. International Centre of Excellence for Wearable Electronics and Smart Products Conference, (ICEWES), December 10–11, 2002, Cottbus, Germany
22. R. Tummala, E. Rymaszewski, A. Klopfenstein: *Microelectronics Packaging Handbook: Semiconductor Packaging* (Kluwer Academic, Dordrecht 1997)

23. G. Kelly, A. Morrissey, J. Alderman, H. Camon: 3D packaging methodologies for microsystems. *IEEE Transactions on Advanced Packaging*, Vol. 23, No. 4, November 2000, pp. 1–8
24. J. Barrett, C. Cahill, T. Compagno, M. O Flaherty, T. Hayes, W. Lawton, J. O Donavan, C. O'Mathuna, G. McCarthy, O. Slattery, F. Waldron: Performance and reliability of a three-dimensional plastic moulded vertical multichip module (MCM-V). *Proceedings of IEEE Electronics Components and Technology Conference*, Las Vegas, May 1995, pp. 656–663
25. B. Majeed, K. Delaney, J. Barton, J. O'Brien, A. Kelleher, S.C. O'Mathuna: Implementing ultra thin autonomous modules for ambient systems applications using 3-D packaging techniques. *Proceedings of the IMAP Nordic Conference*, 21–24 September 2003, Helsinki University of Technology, Espoo, Finland
26. A. Mathewson, J. Alderman: The snake—a novel high area utilisation approach for creating integrated circuits in fibre form. *Irish Preliminary Patent P129447*
27. S.F. Al-Sarawi, D. Abbott, P. Franzon: Review of 3D VLSI packaging technology. *IEEE Transactions on Components, Packaging, and Manufacturing Technology Part B: Advanced Packaging*, vol. 21, No. 1, February 2002, pp. 2–14
28. K. Delaney: Systems packaging issues for ambient intelligence systems platforms. *IMAPS Nordic Tutorial*, 21 September 2003, Helsinki University of Technology, Espoo, Finland
29. J. Lau, R. Lee: *Chip Scale Package, (CSP): Design, Materials, Processes, Reliability, and Applications* (McGraw-Hill, New York 1999)
30. Tessera's approach to stacked IC's packaging, www.Tessera.com
31. P. Clot, J.-F. Zeberl, J.-M. Chenuz, F. Ferrando, D. Styble: Flip-chip on flex for 3D packaging, *Proceedings of the IEEE/CPMT Electronics Manufacturing Technology Symposium*, Austin, Texas, October 18–19, 1999, pp. 36–41
32. Irvine Sensors Corporation, Neo stacking technology, www.Irvine-Sensors.com
33. 3D-Plus website, www.3d-plus.com
34. M. Sunohara, T. Fujii, M. Hosino, H. Yonemura, M. Tomisaka, K. Takahashi: Development of wafer-thinning and double-sided bumping technologies for 3-D stacked LSI. *Proceeding of the IEEE 52nd Electronics Component and Technology Conference (ECTC 2002)*, San Diego, California, May 28–31, 2002, pp. 238–245
35. T. Kenji, T. Hiroshi, T. Yoshihiro: Current status of research and development of three dimensional chip stack technology. *Japanese Journal of Applied Physics* **40**, 3032 (2001)
36. Y. Yano, T. Sugiyama, S. Ishihara, Y. Fukui, H. Juso: Three-dimensional very thin stacked packaging technology for SiP. *Proceedings of the IEEE 52nd Electronics Component and Technology Conference ECTC 2002*, May 28–31, 2002, pp. 1329–1334
37. J. Barton, K. Delaney, S. Bellis, S.C. O'Mathuna, J.A. Paradiso, A. Benbasat: Development of distributed sensing systems of autonomous micro-modules. *53rd Electronic Components and Technology Conference ECTC 2003*, May 27–30, 2003, New Orleans, USA
38. F. Gemperle, C. Kasabach, J. Stivoric, M. Bauer, R. Martin: Design for wearability. *Proceedings of the Second International Symposium on Wearable Computers*, Pittsburgh, PA, October 1998
39. The fibre computing project (FiCom), www.fibercomputing.net

40. T. Lisby: Mechanical characterisation of flexible silicon microstructures. Proceedings of the 14th European Conference on Solid-State Transducers, August 2000, Copenhagen, Denmark, pp. 279–281
41. F. Ericson, J. Schweitz: Micromechanical fracture strength of silicon. *Journal of Applied Physics* **68**, 5840 (1990)
42. T. Healy, A. Mathewson, J. Alderman, J. Donnelly: Development of a novel technology for building flexible and wearable integrated systems. 53rd Electronic Components and Technology Conference ECTC 2003, May 27–30, 2003, New Orleans, USA
43. J. Kou, K. Su: *CMOS VLSI Engineering Silicon on Insulator (SOI)* (Kluwer Academic, Boston 1998) pp. 15–59
44. K. Harsh: Dealing with MEMS stiction and other sticking problems (University of Colorado), <http://mems.colorado.edu/c1.gen.intro/c2.notes/stiction.sht>
45. T. Stieglitz, H. Beutel, M. Schuettler, J.-U. Meyer: *Micromachined, Polyimide-Based Devices for Flexible Neural Interfaces* (Kluwer Academic, Dordrecht 2000) pp. 283–284

25 Semiconductors with Brain*

P. Fromherz

25.1 Introduction

Both computers and brains work electrically. But their charge carriers are different—electrons in a solid ion lattice and ions in a polar fluid. It is an intellectual and technological challenge to join these different systems directly on the level of electronic and ionic signals [1, 2]. Already in the 18th century, Luigi Galvani established the electrical coupling of inorganic solids and excitable living tissue. Today, after 50 years of dramatic developments in semiconductor microtechnology and cellular neurobiology, we may envisage such an integration by far more complex interactions, right on the level of individual nerve cells and microelectronic devices. In a first step, however, we have to elucidate the fundamental biophysical mechanisms of bioelectronic interfacing on the scale of nanometers, micrometers and millimeters. If we succeed in this endeavour, we shall be able to fabricate iono-electronic devices to solve problems in molecular biology, to develop neuroelectronic assemblies to study the physics of brain-like systems, and to contribute to medicine and information technology by creating microelectronic neuroprostheses and nerve-based ionic processors.

25.2 Iono-electronic Interfacing

A hybrid chip with a neuron from rat brain and with transistors in silicon is shown in Fig. 25.1. A nerve cell (diameter about $20\text{ }\mu\text{m}$) is surrounded by a membrane with an electrically insulating core of lipid. This lipid bilayer (thickness about 5 nm) separates the environment with about 150 mM (10^{20} cm^{-3}) sodium chloride from the intracellular cytoplasm with about 150 mM potassium chloride. Ion current through the membrane is mediated by protein molecules, the ion channels. Silicon is suitable as an electronically conductive substrate for three reasons: (i) Coated with thermally grown silicon dioxide (thickness $10 - 1000\text{ nm}$), silicon is a perfect inert substrate for

* Reprinted from PHYSICA E, **16**, Peter Fromherz, Semiconductor chips with ion channels, nerve cells and brain, pp. 24-34, Copyright 2003, with permission from Elsevier.

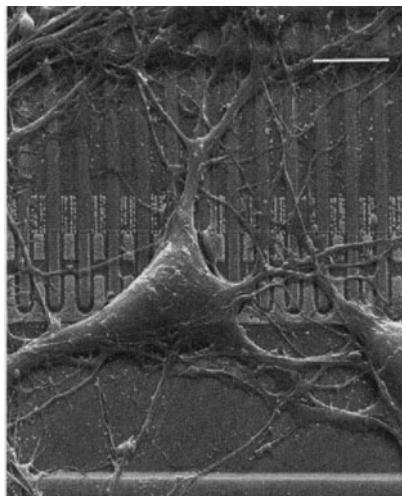


Fig. 25.1. Nerve cell from rat brain on silicon chip. Electron micrograph, scale bar 10 μm . The surface of the chip consists of thermally grown silicon dioxide. The metal-free gates of a linear array of field-effect transistors are visible as dark squares. The neuron is cultured on the chip for several days in an electrolyte

culturing nerve cells. (ii) The thermally grown silicon dioxide suppresses the transfer of electrons and concomitant electrochemical processes that lead to corrosion of silicon and damage of the cells. (iii) An established semiconductor technology allows the fabrication of microscopic devices in direct contact to the cells.

In principle, a coupling of ionic signals in a neuron and electronic signals in the semiconductor can be attained by electrical polarization. If the insulating lipid bilayer is in direct contact with the insulating silicon dioxide, a compact dielectric is formed. An electrical field across the membrane, as created by neuronal activity, polarizes the oxide such that the electronic band structure of silicon is affected. Vice versa, an electrical field across the oxide, as caused by a voltage applied to the chip, polarizes the membrane such that the conformation of membrane proteins is affected. However, when a nerve cell grows on a chip, we cannot expect that the lipid and oxide form a compact dielectric layer. Cell adhesion is mediated by protein molecules that protrude from the membrane and that are deposited on the substrate. These proteins keep the lipid core of the membrane at a certain distance from the substrate, stabilizing a cleft between cell and chip that is filled with electrolyte. That conductive cleft suppresses a mutual polarization of silicon dioxide and cell membrane.

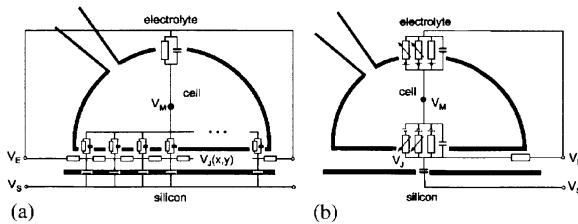


Fig. 25.2. Core-coat conductor of cell-semiconductor junction. The heavy lines indicate silicon dioxide, cell membrane and micropipette. The cross sections are not to scale: the distance of membrane and chip is 10 – 100 nm, the diameter of a cell is 10 – 100 μm . (a) AC circuit of area-contact model. The infinitesimal elements of oxide, membrane and electrolyte film in the junction are represented as capacitors and ohmic resistances. (b) DC circuit of point-contact model with voltage-gated ion conductances. Oxide, membrane and electrolyte film are represented by global capacitances and resistances. V_M , V_s and V_E are the electrical potentials in the cell, in the chip and in the bath. V_j is the transductive extracellular potential (TEP) in the junction

25.2.1 Planar Core-Coat Conductor

A cell-silicon junction forms a planar electrical core-coat conductor: membrane and silicon dioxide insulate a conductive cleft from cytoplasm and silicon [3]. To describe current and voltage, we use the two-dimensional area-contact model or the zero-dimensional point-contact model as sketched in Fig. 25.2. Electrical signals in the cell induce ionic and displacement currents through the membrane. The concomitant current along the cleft gives rise to a transductive extracellular potential (TEP) between the cell and chip. The resulting electrical field across the oxide can be probed by a field-effect transistor. On the other hand, a voltage transient applied to the chip leads to a displacement current through the oxide. Again a TEP appears between chip and cell related with the concomitant current along the cleft. This TEP induces an electrical field across the membrane that can be probed by voltage-sensitive ion channels. The cell-chip interaction through the core-coat conductor is promoted (i) by a high resistance of the core, i.e. a small width d_J and a high specific resistance ρ_J of the cleft, and (ii) by a large current through the coats, i.e. by large area-specific ion conductances g_{JM}^i in the attached membrane and a large area-specific capacitance c_S of the oxide.

25.2.2 Cleft of Cell-Silicon Junction

Silicon reflects light such that standing modes of the electromagnetic field are formed with a node of the electric field near the surface. We take advantage of this effect to measure the distance of cells and silicon using fluorescent dye molecules as electromagnetic antennas (Fluorescence Interference Contrast

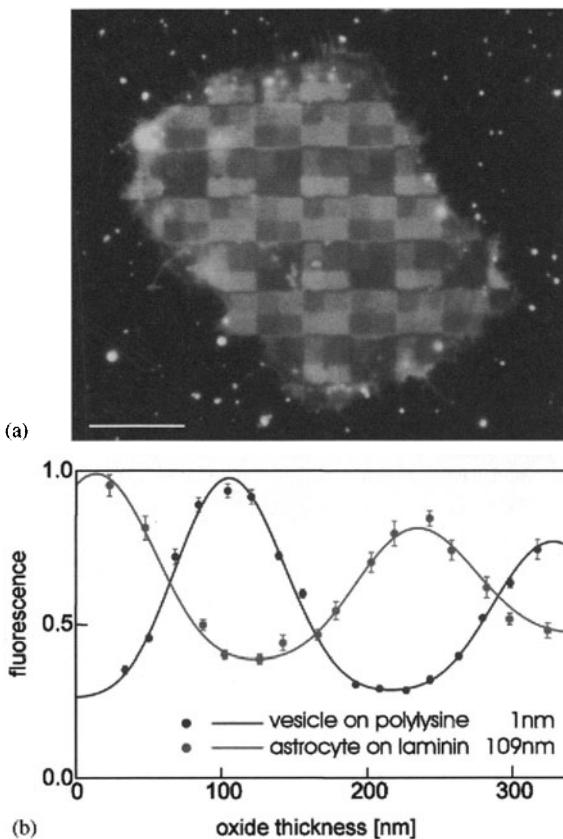


Fig. 25.3. Fluorescence interference contrast (FLIC) microscopy. (a) Fluorescence micrograph of the adhesion region of a rat astrocyte on a silicon chip with quadratic $2.5 \mu\text{m} \times 2.5 \mu\text{m}$ terraces of silicon dioxide. Scale bar $10 \mu\text{m}$. The chip is coated with laminin. The membrane is stained with a cyanine dye. (b) Fluorescence intensity versus height of the terraces for the astrocyte (bright dots) and for a lipid vesicle on polylysine (dark dots). The lines are computed by an electromagnetic theory with a water film of 109 nm between oxide and membrane for the astrocyte and of 1 nm for the lipid vesicle

Microscopy) [4, 5]. We fabricate silicon chips with microscopic oxide terraces (size $2.5 \mu\text{m} \times 2.5 \mu\text{m}$, step height about 20 nm), culture neuronal cells and label the lipid core of the membrane with a fluorescent dye. The terraces of defined height together with the cleft of unknown width place the membrane at different positions in the standing modes. As a consequence, light absorption and light emission of the dye are modulated.

The fluorescence of a cell membrane on silicon with oxide terraces is shown in Fig. 25.3a for an astrocyte from rat brain on a chip coated with laminin. The checkerboard pattern matches the oxide terraces. The homogeneous in-

tensity on each terrace and the repetitive pattern on the unit cells of 4×4 terraces indicate that a well-defined distance between membrane and chip exists. The intensity is plotted in Fig. 25.3b versus the height of the terraces. Surprisingly, it has a maximum on the thinnest oxide, drops then and increases again. For comparison, we attach a pure lipid bilayer using polylysine. There the intensity starts with a minimum on the thinnest oxide. When we fit the data with the electromagnetic theory of dipole radiation [4] we find that the pure bilayer is in direct contact to the oxide (distance 1 nm), whereas a cleft of 109 nm exists between the chip and the lipid core of the cell membrane [5]. That separation is caused by proteins in the membrane (glycocalix) and on the chip (laminin). Choosing different neuronal cells and different coatings, we could not achieve contacts closer than 40 nm. It is an important task to reduce that distance by physical and chemical modifications of the chip and by genetic modifications of the membrane, without impairing the viability of the cells.

25.2.3 Conductance of the Cleft

When an AC voltage $\underline{V}_M(\omega)$ of angular frequency ω is applied to a cell with the bath on ground potential, a voltage $\underline{V}_J(\omega)$ is induced in the junction (Fig. 25.2). From the complex spectral transfer function $\underline{V}_J/\underline{V}_M$ we obtain the sheet resistance r_J of the cleft on the basis of the area- or point-contact model [3, 6]. The voltage \underline{V}_M is applied by a micropipette, the voltage \underline{V}_J is measured with an open field-effect transistor in the substrate. A transistor with leech neuron and micropipette is depicted in Fig. 25.4a. Amplitude and phase of the transfer function are plotted in Figs. 25.4b and c versus the frequency $f = \omega/2\pi$. We observe two types of spectra [3]: an A-spectrum with a small amplitude at low frequencies, an increase of the phase at 10 Hz and of the amplitude above 1000 Hz. A B-spectrum with a high amplitude at low frequencies.

We evaluate the experiment with the point-contact model with given area-specific capacitances $c_M = 5 \mu\text{F}/\text{cm}^2$ and $c_S = 0.3 \mu\text{F}/\text{cm}^2$ of membrane and chip and with unknown area-specific ohmic conductances g_{JM} and g_J of attached membrane and cleft. The low-frequency limit of the amplitude is determined by the ohmic conductances. Thus, the small amplitude of the A-spectrum indicates a low membrane conductance g_{JM} . The phase increases when the capacitive conductance of the membrane overcomes the ohmic conductance of the membrane, the amplitude increases when it overcomes the ohmic conductance of the cleft. When we fit the data we obtain $g_{JM} \approx 0.35 \text{ mS}/\text{cm}^2$ and $g_J \approx 210 \text{ mS}/\text{cm}^2$. The cleft conductance is by a factor 600 higher than the membrane conductance of the intact cell. In the B-spectrum, the situation is different. From a fit of the data we obtain $g_{JM} \approx 40 \text{ mS}/\text{cm}^2$ and $g_J \approx 40 \text{ mS}/\text{cm}^2$, i.e. the cleft conductance is lowered by a factor of five, whereas the membrane conductance is enhanced by two orders of magnitude. The B-spectrum reflects a damaged membrane. Using

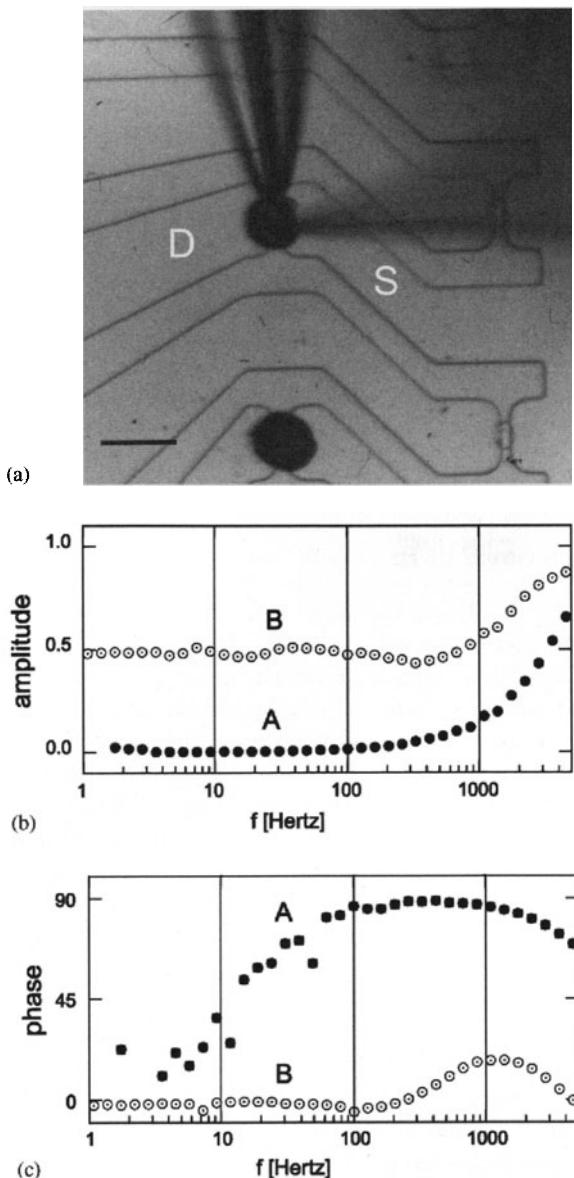


Fig. 25.4. Voltage transfer function V_J/V_M from cell to cell-silicon junction. (a) Micrograph of leech neuron on open field-effect transistor with source S and drain D. Scale bar 100 μ m. The cell is contacted by a patch micropipette. From the right, a second pipette is impaled to measure the actual voltage in the cell. (b) Amplitude of voltage transfer V_J/V_M from cell to junction versus frequency f . (c) Phase of voltage transfer. The dots mark an A-type spectrum, the circles a B-type spectrum

the relation $g_J = 5/r_J a_J^2$ between the area-specific conductance g_J , the sheet resistance r_J and the radius a_J of the cleft, we obtain for the A-type contact, with an area $a_J^2 \pi = 1000 \mu\text{m}^2$, a sheet resistance $r_J \approx 8 \text{ M}\Omega$. It is a challenge to enhance the sheet resistance of the cleft by reducing its width or by enhancing its specific resistance.

25.2.4 Ion Channels in Cell-Silicon Junction

The transductive extracellular potential $V_J(t)$ during neuronal excitation depends on the current through ion conductances in the attached membrane. On the other hand, ion conductances in the attached membrane are the primary target of the transductive extracellular potential $V_J(t)$ induced by the capacitive stimulation. Thus, we have to ask: Are there functional ion channels in the contact region at all? Is the density of ion channels the same as in the free membrane [7]? A detailed characterization is achieved with well-defined recombinant channels [8].

Human embryonic kidney cells (HEK293) with overexpressed hSlo potassium channels on a transistor array are depicted in Fig. 25.5a. A DC voltage V_M is applied to a cell on a transistor using a micropipette, the resulting current I_M through the total membrane and the extracellular voltage V_J in the junction are simultaneously measured. Positive voltages enhance the pipette current and the transistor signal as shown in Fig. 25.5b. Plotting V_J versus I_M , we find a perfect linear correlation [8]. The current through the total membrane area A_M is given by $I_M A_M \approx g_M^K (V_M - V_0^K)$ with the average specific conductance g_M^K and the reversal voltage V_0^K . The extracellular voltage is described by $g_J V_J \approx g_{JM}^K (V_M - V_0^K)$ with the specific conductance g_{JM}^K in the contact. The proportionality of V_J and I_M indicates $g_{JM}^K / g_M^K = \text{const}$. We conclude: The local conductance g_{JM}^K and the global conductance g_M^K obey an identical relation of voltage-dependent gating. Functional potassium channels exist in the contact. With scaling factors A_M and g_J obtained from AC measurements, we obtain a ratio $\bar{g}_{JM}^K / \bar{g}_M^K = 3.3$ of the maximum conductances, indicating that the hSlo potassium channels are significantly accumulated in the junction. An optimization of this sorting of ion channels into a neuron-chip contact is an important future task.

25.3 Neuron-Silicon Circuits

The first step towards an integration of neuronal dynamics and digital electronics is an interfacing of individual nerve cells and silicon microstructures: (i) eliciting neuronal activity from the chip by capacitive stimulation and, (ii) recording neuronal activity by a transistor. In a next stage, pairs of nerve cells are coupled to a chip with two fundamental pathways: (i) a signaling neuron-silicon-silicon-neuron with recording the activity of one neuron by a transistor, signal transfer through the microelectronics of the chip and

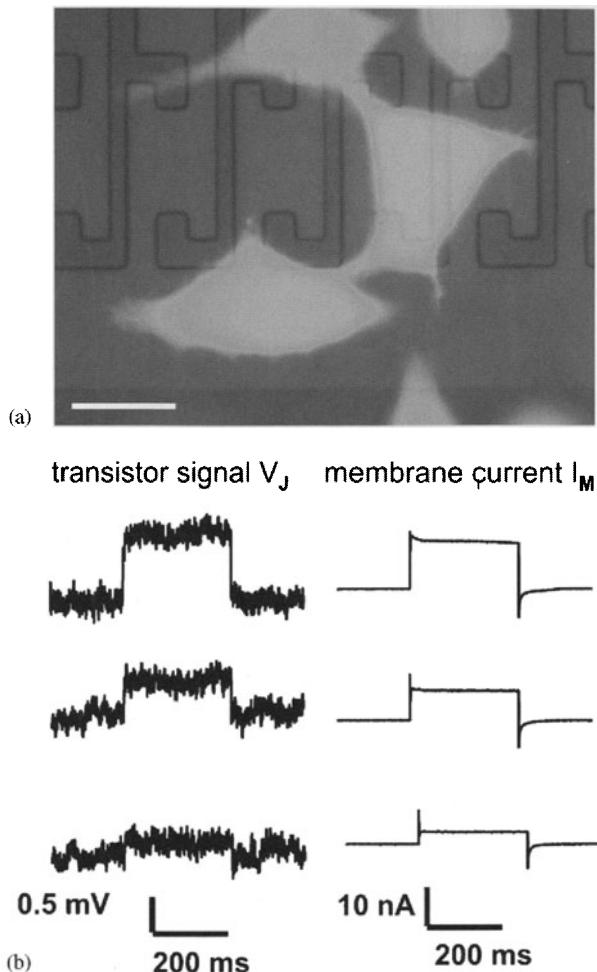


Fig. 25.5. Recombinant hSlo potassium channels on transistor. (a) Transfected HEK293 cells (*bright*) on linear transistor array (*dark*). (b) Iono-electronic coupling at three intracellular voltages $V_M = 30, 45$ and 58 mV . (*Left*) Extracellular voltage V_J in the junction. (*Right*) Current I_M through the total cell membrane

capacitive stimulation of a second neuron, and (ii) a signaling silicon-neuron-neuron-silicon with capacitive stimulation of one neuron, signal transfer through a synapse to a second neuron and transistor recording of neuronal activity there.

25.3.1 Transistor Recording of Neuronal Activity

An action potential consists in an opening of sodium channels with a current into the cell and a delayed opening of potassium channels with a compen-

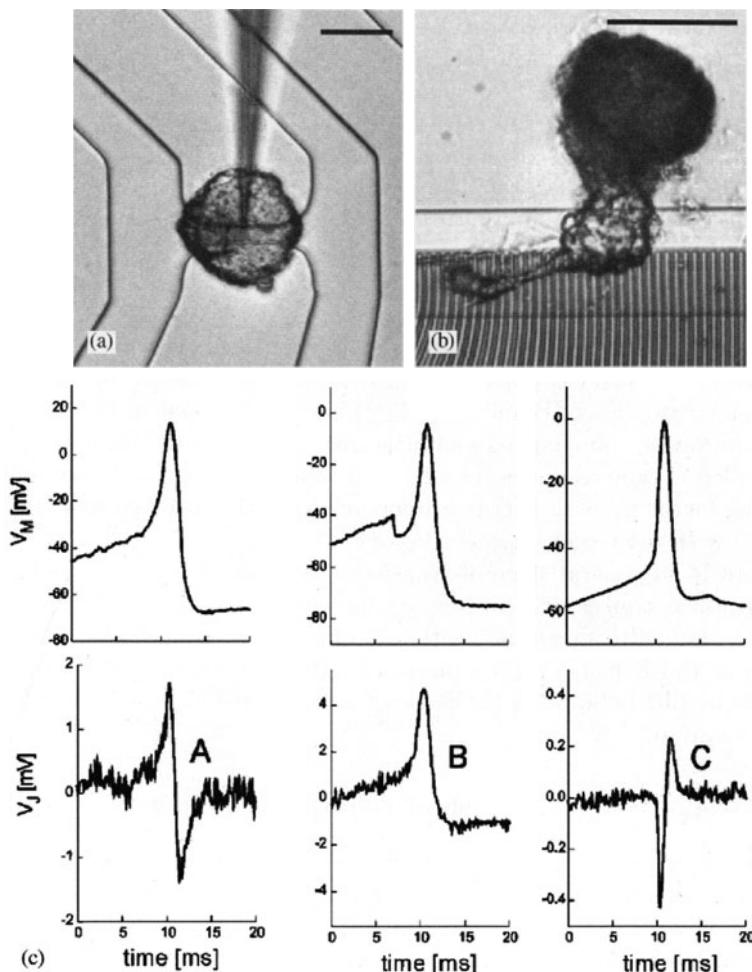


Fig. 25.6. Transistor recording of neuronal excitation. (a) Cell body of a leech neuron on an open field-effect transistor. Scale bar 50 μ m. The cell is impaled with a micropipette. (b) Axon stump of a leech neuron on an array of field-effect transistors. Scale bar 50 μ m. (c) A-, B- and C-type coupling. The upper row shows the intracellular voltage $V_M(t)$, the lower row the extracellular voltage $V_J(t)$ in the junction. A and B-type couplings are observed for arrangement (a), C-type coupling for arrangement (b)

sating outward current. It drives ionic and capacitive current through the membrane attached to a chip. This current is squeezed through the cleft and gives rise to a transductive extracellular potential $V_J(t)$ that is probed by a transistor [9–11].

Transistor records of leech neurons are shown in Fig. 25.6. Two geometries are depicted in Figs. 25.6a and b with the cell body on a transistor and

with the axon stump on a transistor array. The neurons are impaled with a micropipette and action potentials are elicited by current injection. The intracellular potential $V_M(t)$ is measured with the pipette. The response of the transistor is calibrated in terms of the transductive extracellular potential $V_J(t)$. Three records are depicted in Fig. 25.6c: (i) With a cell body on a transistor, $V_J(t)$ resembles the first derivative of $V_M(t)$. (ii) With a cell body on a transistor, $V_J(t)$ resembles the $V_M(t)$ itself. (iii) With an axon stump on a transistor, $V_J(t)$ resembles the inverted first derivative of $V_M(t)$.

If the attached membrane contains no voltage-gated conductances, the current balance in the junction is given by $g_J V_J \approx g_{JM} V_M + c_M d V_M / d t$. For a negligible leak conductance g_{JM} (A-junction), the capacitive current dominates, and $V_J(t)$ is proportional to the first derivative of $V_M(t)$. For a high leak conductance (B-junction), $V_J(t)$ is proportional to $V_M(t)$. Selective accumulation of voltage-gated channels may give rise to various waveforms $V_J(t)$ called C-type responses [11,12]. If all conductances are accumulated by the same factor $\mu_J > 1$, $V_J(t)$ is proportional to the inverted first derivative of $V_M(t)$ with $g_J V_J \approx (1 - \mu_J) c_M d V_M / d t$ [11, 12].

There is no general shape of transistor records of action potentials. We may optimize transistor recording (i) by improving the cell-chip contact, reducing the width or enhancing the specific resistance of the cleft, (ii) by enhancing the density of ion channels in the junction using recombinant methods or (iii) by lowering the noise of the transistors by improved design and fabrication.

25.3.2 Capacitive Stimulation of Neuronal Activity

When a changing voltage $V_S(t)$ is applied to the stimulation area of a silicon chip, displacement current flows through the oxide (Fig. 25.2). The concomitant current along the cleft gives rise to a transductive extracellular potential $V_J(t)$, which may open voltage-gated ion channels such that an action potential is elicited [13, 14].

A snail neuron on a two-way contact of a stimulation area and a transistor is depicted in Fig. 25.7a. A burst of three voltage pulses is applied to the stimulator. The intracellular voltage $V_M(t)$ responds with short capacitive transients at the rising and falling edge of each pulse and with an increasing stationary depolarization that leads to an action potential, as shown in Fig. 25.7b [14]. The transistor allows us a look into the junction. The rising and falling edges of each pulse are related with fast capacitive transients $V_J(t)$. In addition, at each rising edge a negative transient $V_J(t)$ is induced that slowly decays during the pulse and during the subsequent pulse interval.

When a voltage step of height V_S^0 is applied to a chip, the extracellular and intracellular voltage are exponentials with $V_J \propto \exp(-t/\tau_J)$ and $V_M = V_J \beta_M / (1 + \beta_M)$, where β_M is the ratio of attached and free membrane area, if the membrane conductance is neglected in the initial phase of stimulation. For snail neurons the time constant is around 100 μ s. These primary responses are

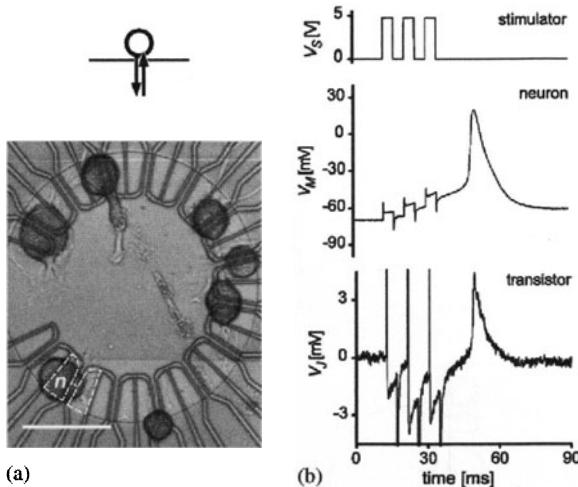


Fig. 25.7. Capacitive stimulation of neuron from silicon chip. (a) Micrograph of snail neurons on a chip with a circular arrangement of two-way contacts. The stimulation area with two wings under neuron n is marked with a dashed line, the transistor is located between the two wings. Scale bar $100\text{ }\mu\text{m}$. (b) (Top) Voltage $V_S(t)$ applied to the stimulation area; (center) intracellular voltage $V_M(t)$ measured with impaled pipette; (bottom) extracellular voltage $V_J(t)$ measured with the transistor

reflected in the records of Fig. 25.7b. The existence of fast voltage transients $V_M - V_J$ in the attached membrane can be directly revealed with voltage-sensitive dyes [15]. The transistor record of Fig. 25.7b indicates that the positive capacitive transients $V_J(t)$ induce an inward current through the attached membrane which decays slowly and is not affected by the negative capacitive transients. This inward current seems to be responsible for the increasing intracellular depolarization $V_M(t)$ that leads to excitation.

A complete rationalization, however, of capacitive stimulation is not attained. Current injection through a leaky membrane, capacitive gating of ion channels and capacitive electroporation are difficult to distinguish. The stimulation may be optimized (i) by lowering the conductance of the cleft, (ii) by inserting recombinant ion channels in the junction and (iii) by stimulation contacts with a higher area-specific capacitance.

25.3.3 Circuits with Two Neurons on Silicon Chip

The hybrid pathway neuron-silicon-silicon-neuron is implemented with two snail neurons attached to two-way contacts of a silicon chip as shown in Fig. 25.8a. The signaling is illustrated by the circuit of Fig. 25.8b: (i) The activity of the first neuron is recorded by a transistor. (ii) The transistor record of an action potential is identified on the chip with a Schmitt trigger.

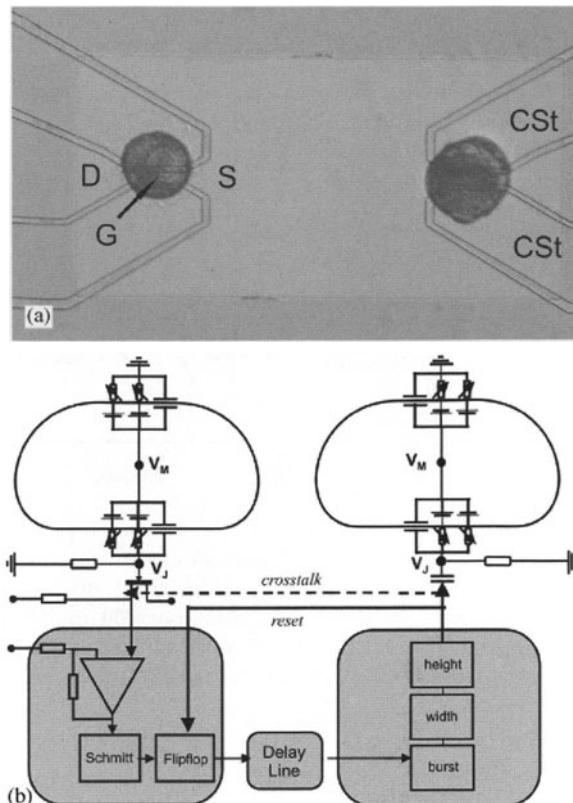


Fig. 25.8. Signaling neuron-silicon-silicon-neuron. (a) Two snail neurons attached to two-way contacts of a silicon chip with source (S), drain (D) and gate area (G) of a transistor and with two wings of a capacitive stimulation area (CSt). (b) Circuit (*Upper part*). Point-contact model of transistor recording and capacitive stimulation. (*Lower part*) Block diagram of microelectronics with recognition (Schmitt trigger), delay line and pulse generator. The dashed line marks the cross talk from stimulator to transistor. The flip-flop is set by an action potential and reset after the stimulation pulses

(iii) The resulting digital signal triggers a delay line. (iv) The delay line triggers a burst of voltage pulses that are applied to a stimulation contact. (v) Neuronal activity in a second neuron is elicited by capacitive coupling. As a result, the second neuron fires in strict correlation to the first neuron, not coupled by a neuronal connection but by a silicon “prosthesis” [16]. The crucial problem of the device is the crosstalk from stimulator to transistor on the chip. The resulting artifacts of the transistor record are eliminated by setting a flip-flop after identification of an action potential and by resetting it after the stimulation pulses. In a first implementation, the chip consists of two

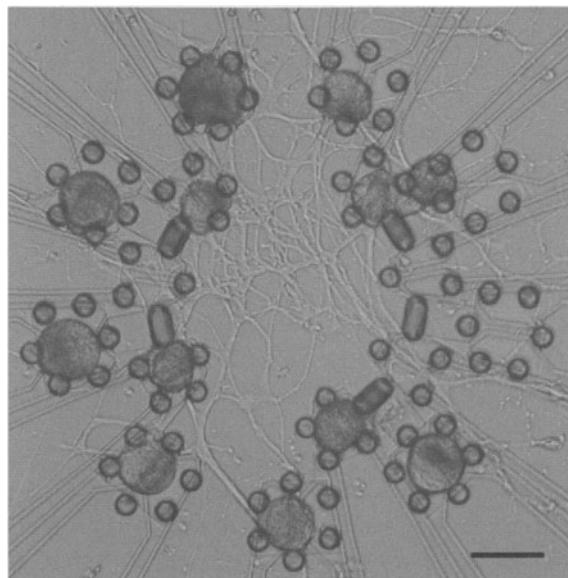


Fig. 25.9. Network of snail neurons on silicon chip. The cell bodies (*dark blobs*) are immobilized by picket fences (*small dots*) on a circle of two-way contacts with neurites grown in the central area (*bright threads*) after 2 days in culture. Scale bar 100 μ m

parts that are wire-bonded side by side, an interface unit with the transistors and stimulation spots and an interneuron unit with the microelectronics.

In a hybrid pathway silicon-neuron-neuron-silicon, two neurons on two-way contacts are connected by grown neurites and a synapse. During out-growth, however, the cell bodies are displaced on the chip such that the junctions with transistors and stimulators are disrupted. To overcome the problem, picket fences are fabricated and the neuronal cell bodies are mounted as shown in Fig. 25.9. The cell bodies are kept on the two-way contacts, even after culturing them for several days [17]. The signaling is documented in Fig. 25.10. A burst of voltage pulses is applied to excite neuron 1 as checked with an impaled micropipette. In neuron 2 a subthreshold depolarization is observed that is mediated by an electrical synapse. A second burst of voltage pulses elicits another action potential in neuron 1 and leads to a further depolarization of neuron 2. After the third burst, which fails to stimulate neuron 1, the fourth burst giving rise to an action potential in neuron 1 triggers the excitation of neuron 2. The postsynaptic action potential is recorded by a transistor. Delay and temporal summation of the postsynaptic signals correspond to the intracellular presynaptic stimulation.

Complex neuronal nets rely on (i) a mapping between sets of neurons and (ii) Hebbian learning with an enhanced synaptic strength by correlated presynaptic and postsynaptic activities. Studies of network dynamics require

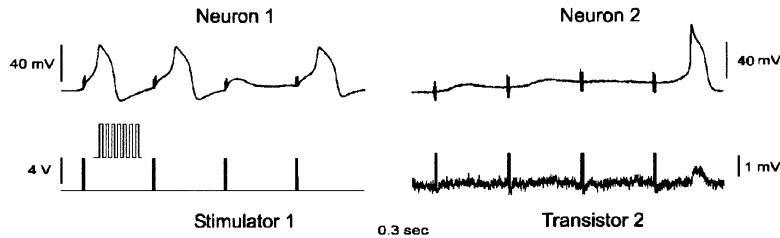


Fig. 25.10. Signaling silicon-neuron-neuron-silicon with electrical synapse between two snail neurons. (*Left bottom*) Bursts of voltage pulses applied to the stimulator (seven pulses, amplitude 5 V, duration 0.5 ms); (*left top*) intracellular voltage of neuron 1; (*right top*) intracellular voltage of neuron 2; (*right bottom*) transistor record of neuron 2

(i) neuronal maps with a defined topology of synaptic connections, and (ii) noninvasive supervision of numerous neurons to induce learning and observe the performance. To achieve these goals, we must control outgrowth and synapse formation [18, 19] and fabricate silicon chips with numerous closely packed arrays of two-way contacts.

25.4 Brain-Silicon Chips

Culturing defined neuronal nets is avoided when we use neuronal nets given by brains. Planar networks are preferred in order to attain an adequate supervision by a planar chip. Organotypic brain slices are promising as they are only a few cell layers thick and conserve major neuronal connections. However, an electronic interfacing of an individual neuron can hardly be achieved. We have to consider a stimulation and recording of local populations of neurons. The concept of a core-coat conductor of an individual cell-chip junction is not adequate.

25.4.1 Tissue-Sheet Conductor

In an organotypic brain slice, the neurons are embedded in a tissue of about 100 μm thickness between the insulating silicon dioxide of a chip and an electrolytic bath on ground potential, as illustrated in Fig. 25.11a. Neurons are local sources or sinks of current that flow to adjacent regions of the tissue layer and to the bath. As a consequence, an extracellular field potential appears in the tissue that may be recorded by transistors in the substrate. On the other hand, capacitive contacts in the substrate may locally inject current into the tissue layer that flows to adjacent regions of the slice and to the bath [20]. The resulting extracellular field potential may elicit neuronal excitation. A brain slice between chip and bath is a sheet conductor

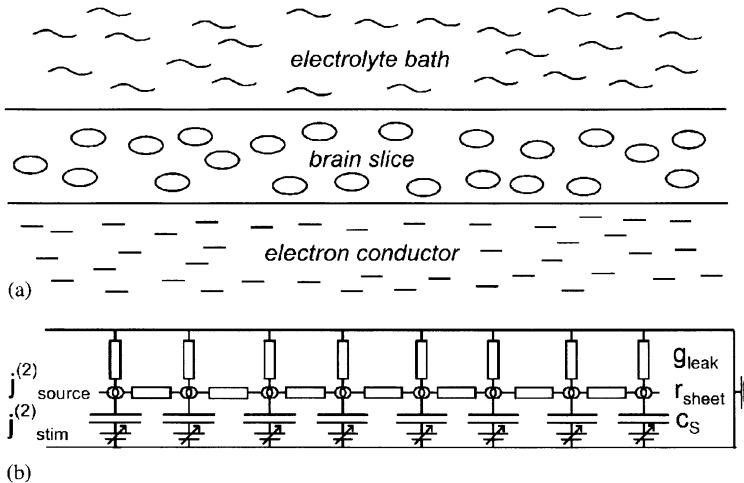


Fig. 25.11. Brain slice on silicon substrate: (a) Geometry of tissue layer between electron conductor and electrolyte bath. (b) Sheet conductor model. The neurons in the slice give rise to a current-source density per unit area. The current flows along the slice (sheet resistance r_{sheet}) and to the bath (leak conductance g_{leak}). The slice is stimulated by current from the chip by capacitive contacts (specific capacitance c_s)

with a capacitive bottom and a leaky cover. We describe the substrate by a capacitance per unit area c_s , the sheet by a resistance r_{sheet} and the leak by an ohmic conductance per unit area g_{leak} , as illustrated by the circuit of Fig. 25.11b. The current source density per unit area $j_{\text{source}}^{(2)}$ and the stimulation current due to a changing voltage V_s applied to the capacitive contacts are balanced by the current along the sheet and by the ohmic and capacitive shunting to the bath and the substrate. A typical feature is the length constant $\lambda_{\text{sheet}} = 1/\sqrt{r_{\text{sheet}} g_{\text{leak}}}$ that determines the range of electrical interactions. The sheet conductor model describes (i) the field potential V_{field} that arises from neuronal activity, as it may be recorded with transistors, and (ii) the field potential that is caused by capacitive stimulation, as it may elicit neuronal excitation.

25.4.2 Transistor Recording of Brain Slice

A slice from rat hippocampus on a silicon chip with a linear array of transistors is shown in Fig. 25.12a. It is stimulated with a tungsten electrode. A profile of evoked field potentials is recorded across the CA1 region [21]. The transistor array is able to detect simultaneously the negative voltage transients in the region of the dendrites (stratum radiatum) where current flows into the neurons and the positive transients at the cell bodies (stratum

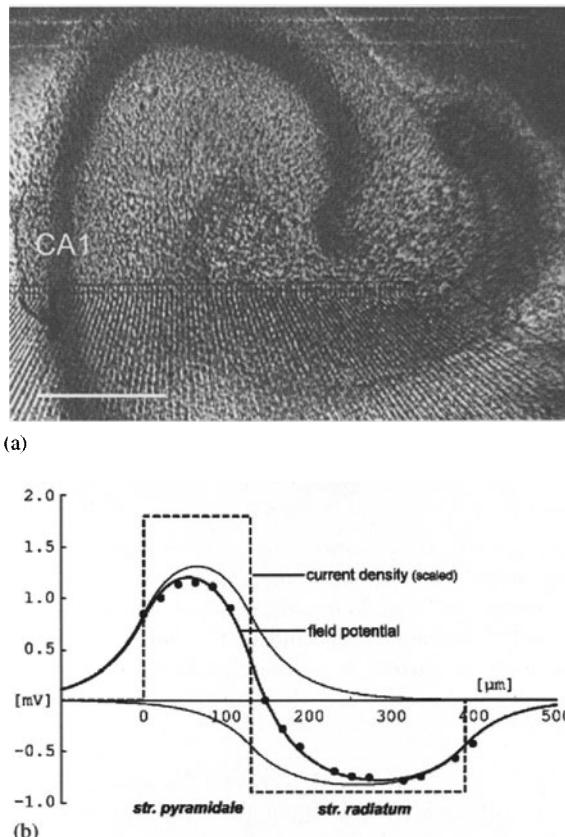


Fig. 25.12. Transistor recording of organotypic slice. (a) Nissl staining of a slice from rat hippocampus cultured for 14 days. Scale bar 400 μm . Neuronal cell bodies (dots) are arranged in bands. A linear transistor array is aligned perpendicular to the CA1 region through the dark band of cell bodies (stratum pyramidale) and the light region with dendrites (stratum radiatum). (b) Amplitude profile of field potentials across the CA1 region (black dots). The data are fitted by the sheet conductor model with a constant negative current density in the stratum radiatum and a balancing constant positive current density in the stratum pyramidale. The contributions from the two strata are drawn as thin lines. The scaled profile of current density is indicated (dashed line)

pyramidal) with a compensating outward current. The amplitudes of the transients are plotted versus the position of the transistors in Fig. 25.12b.

We compare the experimental profile with the sheet-conductor model. Along the CA1 layer the electrical activity of the hippocampus is constant. Across the CA1 layer we assume the simplest current-source density profile that is physiologically meaningful: a constant synaptic inward current density in the stratum radiatum and a compensating constant outward current den-

sity in the stratum pyramidale. The resulting field potential can be expressed by hyperbolic functions. A fit of the data is shown in Fig. 25.12b with a length constant $\lambda_{\text{sheet}} = 50 \mu\text{m}$ and a scaled current density $j_{\text{source}}^{(2)}/g_{\text{leak}} = -0.9 \text{ mV}$. These first results of transistor recording are the basis to implement also capacitive stimulation of organotypic brain slices on silicon chips.

25.5 Summary and Outlook

Basic questions on the electrical interfacing of individual nerve cells and semiconductor chips are fairly well answered: the nature of the core-coat conductor, the properties of the cleft, the role of accumulated ion channels, the mechanism of transistor recording and capacitive stimulation. With respect to the latter issue, however, studies on the capacitive gating of ion channels are required. At present, we are faced with problems of optimization. On the side of the neurons, the cell membrane in the junction must be improved with respect to adhesion and conductance using recombinant methods. On the side of the semiconductor, the capacitance of the stimulators must be enhanced and the noise of the transistors must be lowered.

With respect to hybrid systems of neuronal networks and digital microelectronics, we are in an elementary stage. Two directions are envisaged: small defined networks of neurons from invertebrates and mammals may be created with learning synapses and with defined topology. Large neuronal nets may be grown on closely packed arrays of two-way interface contacts. An adaptation of the industrial standard of CMOS technology is crucial. The interfacing of brain slices is in its infancy. Two directions may be envisaged: Arrays of two-way contacts will lead to a complete spatiotemporal mapping of brain dynamics. Hybrid neuroelectronic systems are implemented that exhibit an associative memory under digital control.

The availability of involved integrated neuroelectronic systems will help to unravel the nature of information processing in neuronal networks and will give rise to new and fascinating physical-biological-computational questions. Of course, visionary dreams of bioelectronic neurocomputers and microelectronic neuroprostheses are unavoidable and exciting, but they should not obscure the numerous practical problems.

Acknowledgments

The work reported in this paper was possible only with the cooperation of numerous students who contributed with their skill and enthusiasm. Most valuable was also the help of Bernt Müller (Institut für Mikroelektronik, TU Berlin) who fabricated several chips and gave his advice for our own chip technology. The project was supported by the Universität Ulm, the Fonds der Chemischen Industrie, the Deutsche Forschungsgemeinschaft, the Max-Planck-Gesellschaft and the Bundesministerium für Bildung und Forschung.

References

1. P. Fromherz: 20th Winter seminar on Molecules, Memory and Information, Klosters, Switzerland (1985) (see <http://www.biochem.mpg.de/mnphys/>)
2. P. Fromherz: *Chem. Phys. Chem.* **3**, 276 (2002)
3. R. Weis, P. Fromherz: *Phys. Rev. E* **55**, 877 (1997)
4. A. Lambacher, P. Fromherz: *J. Opt. Soc. Am. B* **19**, 1435 (2002)
5. D. Braun, P. Fromherz: *Phys. Rev. Lett.* **81**, 5241 (1998)
6. R. Weis, B. Müller, P. Fromherz: *Phys. Rev. Lett.* **76**, 327 (1996)
7. S. Vassanelli, P. Fromherz: *J. Neurosci.* **19**, 6767 (1999)
8. B. Straub, E. Meyer, P. Fromherz: *Nature Biotech.* **19**, 121 (2001)
9. P. Fromherz, A. Offenhäusser, T. Vetter, J. Weis: *Science* **252**, 1290 (1991)
10. M. Jenkner, P. Fromherz: *Phys. Rev. Lett.* **79**, 4705 (1997)
11. R. Schätzthauer, P. Fromherz: *Eur. J. Neurosci.* **10**, 1956 (1998)
12. P. Fromherz: *Eur. Biophys. J.* **28**, 254 (1999)
13. P. Fromherz, A. Stett: *Phys. Rev. Lett.* **75**, 1670 (1995)
14. M. Jenkner, B. Müller, P. Fromherz: *Biol. Cybernet.* **84**, 239 (2001)
15. D. Braun, P. Fromherz: *Phys. Rev. Lett.* **86**, 2905 (2001)
16. P. Bonifazi, P. Fromherz: *Adv. Mater.* **14**, 1190 (2002)
17. G. Zeck, P. Fromherz: *Proc. Nat. Acad. Sci. USA* **98**, 10457 (2001)
18. A.A. Prinz, P. Fromherz: *Biol. Cybernet.* **82**, L1 (2000)
19. M. Merz, P. Fromherz: *Adv. Mater.* **14**, 141 (2002)
20. P. Fromherz: *Eur. Biophys. J.* **31**, 228 (2002)
21. B. Besl, P. Fromherz: *Eur. J. Neurosci.* **15**, 999 (2002)

Index

III-V compounds 27
III-V semiconductor 28
3-D packaging 499
 7×7 reconstruction 119, 191

aberrations 406
ab-initio local-density-functional cluster calculations 268
ab initio molecular dynamics for liquids 14
ab initio pseudopotentials 9
absorption 171
absorption edge 126
absorption measurements 183
acceptor 261, 280
accumulation layer 322, 327, 328
action potential 522
activated processes 208
activation 220
activation barrier 95
activation energy 157, 172
activation energy for deposition 54
activation enthalpy 208
active carbon control 440
adatoms 193
adiabatic approximation 19
admittance spectroscopy 172
admittance techniques 182
ADP, atmospheric downstream plasma 494
adsorption energy 95
adsorption process 57
AFM, atomic force microscope 149, 191
AIC, aluminum-induced crystallization 61
Airflow microsensors 418
Alferov 424

alkali-ion-selective membrane 416
AmI, ambient intelligence 489
AMLCDs, active-matrix liquid crystal displays 49
amorphous growth regime 103
amorphous layer 213
amorphous oxides 219
amphoteric behaviour 437
amphoteric impurity 261, 263
amplifying gate 314
amplifying-gate thyristor 314
angular-coherence condition 225
angularly dependent force 203
annealing 125, 144, 207, 212, 233
annealing stages 221
antibonding site 271
antireflection coating 83, 223
antireflective coating 86
antisite atom 436
APCVD, atmospheric-pressure CVD 52
A-process 28
armchair configuration 478
Arrhenius 267, 268, 271
Arrhenius-type plot 157
ASIC, application-specific integrated circuit 365, 399
a-Si:D 133
a-Si:H solar cells 124
astigmatism 406
astrocyte 518
atomic resolution 194
ATP, adenosine triphosphate 418
avalanche effect 301
avalanche multiplication 383
avalanche radiation 231, 233
average coordination number 123
axial channeling 216

back-bonding 271
 backscattered secondary electrons 407
 ballistic 478
 ballistic carrier transport 391
 bandgap 27, 29, 263
 bandgap engineering 7
 band structure 26
 Bardeen 293
 barrier for diffusion 268
 BC, bond-centred site 263, 265
 beam blanker 406
 beam modulation 99
 binding energy 95, 172
 bipolar transistor effect 26
 Birefringence 419
 blistering 154
 Bloch's principle 123
 Bloch states 127
 blocking state 298
 blocking voltage 231
 blocking-voltage capability 295, 297, 300, 311, 333
 Boersch effect 403, 406
 Bohr radius of an electron 465
 Boltzmann 209
 Boltzmann constant 303
 bond breaking 133
 bonding configuration 8, 126
 B-process 28
 Born–Oppenheimer approximation 3
 boron isotope effect 272
 Boule annealing 447
 Brattain 293
 breakdown behaviour 231
 breakdown decreases 345
 breakdown voltage 113, 246
 Bridgman crucibles 445
 Brillouin zone 30, 470
 BSF, back surface field 85
 bubble formation 439
 buffer 343
 buffer layer 311, 416
 bulk minority carrier diffusion length 82
 Burgers vector 100
 C₃N₄ 215
 CAGR, constant average growth rate 372
 capacitance 352
 capacitance–voltage techniques 263
 capture constants 175, 180
 capture cross-sections 174
 capture processes 173
 capture properties 188
 capture rates 172
 Carbon boats 425
 carbon nanotube 392, 394
 carbon nanotube field effect transistor 482
 carbon nanotube transistors 395
 carrier concentration 302, 308
 carrier distribution 329
 carrier effective mass 65
 carrier lifetime 30, 302
 carrier mobility 27, 29, 35
 carrier recombination 431
 cascade process 175
 cast-ingot technologies 74
 catalyst-mediated CVD growth 477
 CCM PFC applications 359
 CCM, continuous mode 358
 CeF, center faulted 194
 cell bodies (stratum pyramidale) 530
 CeU, center unfaulted 194
 CFLs, cold fluorescent lamps 342
 chalcogen deep double donors 274
 chalcogenide glasses 123
 changes in entropy 174
 changes of sites 207, 209
 channel 321, 324
 channel conductivity 473
 channel resistance 328
 channel width 324
 characteristic diffusion length 208
 charge balance 348
 charge compensation degree 348
 charge state 172, 249, 253, 473
 charge storage time 252
 charge transport 246
 charged states 127
 chemical potential 208, 221, 437, 455
 ChemSage code 438
 chip-level stacking 500
 chip-shrinking 335
 chirality 478
 chromatic aberration 406

CIS, Cu-In-Se 87
 Clark-type gas sensor 417
 cleavage plane 9
 cluster 16, 17
 CMOS, complementary metal-oxide-semiconductor 49, 363, 376
 CNTFETs, carbon nanotube field effect transistors 477
 coalescence 57
 CoF, corner faulted 194
 coherent AmI system 494
 coherent superposition 467
 cohesive energy 13
 cohesive-energy calculations 13
 collision cascade 219
 columnar crystal growth 76
 columnar structure 56, 82
 coma 406
 communications infrastructure 490
 commutation behavior 355
 compensation 269
 compensation devices 347
 compensation parabola 352
 compensation principle 346
 complementary NOR gate 483
 concentration gradient 208
 concentration profile 212
 conducting state 298, 302
 conduction power 344
 conductivity
 n-type 207
 p-type 207
 conductive cleft 516
 conformation of membrane proteins 516
 contact printing 43
 contact resistances 481
 contacting wave 151
 continuity equation 208
 conversion efficiencies 246
 Cool-MOSTM 328, 347
 coordination number 126
 copper contamination 245
 copper interconnects 245
 copper-acceptor pairing 248
 COPs, crystal-originated particles 455
 CoU, corner unfaulted 194
 Coulomb
 blockade 394
 energy 392
 interaction 403
 potential 217
 repulsion 250
 covalent bond 7, 15, 193
 covalent bonding 151, 261
 covalent radius 214
 C-process 31, 43
 crack formation 454
 cracks 156
 critical dose 144
 critical thickness 100, 103
 CRN, continuous random network 123
 cross-sectional specimens 225
 crowdion 210
 CRSS, critical resolved shear stress 438
 crucible techniques 441
 crucible-free floating-zone technique 425
 crystal defects 233
 crystal growth 26
 crystallization temperature 57
 cubic zinc blende structure 427
 current filaments 316
 current gain 252, 318
 current tail 334
 cut-off frequency 35, 36
 CVD, chemical vapor deposition 30, 207
 cw, continuous-wave laser annealing 222
 Cz, Czochralski 73
 CZ, Czochralski pulling process 43
 Czochralski 426
 Czochralski method 31, 441
 dangling bond 10, 11, 37, 63, 86, 193, 270
 dangling Si surface bonds 36
 dark capacitance measurement 172
 Darlington configurations 319, 320
 Darlington structure 320
 DAS, dimer-adatom-stacking-fault 193
 DbyT, dicing by thinning 493

DCM, discontinuous mode 358
 DCS, dichlorosilane 54
 de Broglie formula 406
 Debye screening length 65
 decay 233
 decay time constant 177
 decoherence 467
 decoherence time 468, 472
 deep bandgap levels 277
 deep defects 171, 183
 deep donor level 215
 deep level 219, 251
 deep-level centres 263
 deep-level transient spectroscopy 254
 deep states 64, 127, 261
 defect absorption 129
 defect creation mechanism 130
 defect density 124
 defect engineering 446
 defect ionization energy 249
 defect passivation 257
 defect sites 35
 defective layer 149
 degenerate valence bands 379
 degradation 134, 432
 δ -layer 105, 117
 $(\delta n/\delta p)$ structure 113
 dendrites (stratum radiatum) 529
 dendritic WEB process 78
 density functional theory 9, 11, 13
 density of states 478
 density-of-states distribution 123
 density-of-states effective mass 175
 depletion region 189, 252, 351
 deposition parameter 125
 deposition rate 53, 56
 desorption hydrogen 54
 diamond structure 5, 37
 DIBL, drain-induced barrier lowering 111
 dichroism 271
 dielectric 36
 dielectric constant 188
 dielectric function 220
 dielectric Si_3N_4 phase 214
 differential sensing 417
 diffraction 406
 diffusion 207
 diffusion barrier 439
 diffusion coefficients 247
 diffusion-controlled reaction regime 54
 diffusion length 30, 303
 diffusion-limited pairing reaction 248
 diffusion mechanism 247
 diffusion of hydrogen 157
 diffusion voltage 176
 diffusivity 64
 dimer–adatom–stacking-fault 191
 diode 293, 296
 dipole radiation 519
 direct semiconductors 427
 direct tunneling currents 390
 directional crystallization 426
 disappearing computer 495
 dislocation 29, 82, 85, 210, 256, 448
 dislocation core 100
 dislocation density 146, 380
 dislocation-free growth 429
 dissipation-free operation 467
 dislocation half-loop 101
 dislocation loop 219
 dislocation pinning 44
 dislocation step 96
 divacancies 210
 DLTS, deep-level transient spectroscopy 172, 179, 262, 274
 DMD, digital micromirror 493
 DOF, depth of focus 401
 donor 261
 donor–acceptor pairing 273
 donor–acceptor pairs 250
 dopant 56
 dopant concentration 213, 231
 dopant solubility 63
 doping 27, 126
 doping profiles 104
 DOS distribution 129
 DOS, density-of-states 127
 double-gate transistors 386
 double-tip effect 199
 DRAM, dynamic random access memory 140, 366, 388
 drift fields 65
 drift region thickness 348
 drift zone 112

drifting oxide threshold 363
 drive current 484
 DSPE, double solid-phase epitaxy 142
 dual channels 379
 dual-light-source steady-state photocurrent method 182
 DVDs, digital versatile discs 493

 E3' centre 263
 Early voltage 107
 ECRCVD, electron cyclotron resonance chemical vapor deposition 52, 59
 EDA, electronic design automation 367
 EEPROM, electrically erasable and programmable read-only memory 388
 effective channel mobility 65
 effective density of states 173
 effective electron mass 427
 effective-mass approximation 171
 effective-mass theory 261
 effective modular systems 492
 effusion cells 98
 EFG, edge defined film fed grown 79
 Einstein relation 209
 EL2° concentrations 450
 elastic strain 249
 electric field 299
 electric-field-assisted thermal emission 180
 electrical activity 251
 electrical activity of grain boundaries 62
 electrical capacitive transient signals 257
 electrical polarization 516
 electrical resistivity 449
 electrical synapse 527
 electrochemical potential 415
 electroluminescence 113
 electromigration 481
 electron 478
 electron beam annealing 222
 electron beam irradiation 207
 electron beam lithography 403
 electron beam pulling 74
 electron Bohr magneton 469
 electron effective-mass tensor 470

 electron evaporator 98
 electron-hole plasma 329, 342
 electron irradiation 356
 electron mobility 142
 electron occupancy 177
 electron pair 193
 electron spin 465, 468
 electron tunneling time 466
 electronegativity 245, 427
 electronic degeneracy 173
 electronic states 123
 electronic stopping power 217
 electronic structure
 metal atoms 199
 electronic transport 64
 ELO, epitaxial lateral overgrowth 142
 ELOG, epitaxial lateral overgrowth 432
 Eltran® (epitaxial layer transfer) 142, 159
 EMA, effective-mass approximation 184
 EMC, electromagnetic casting 77
 emission rates 175
 emitter diffusion 49
 emitter efficiency 329
 emitter shorts 314
 empty defect states 177
 encapsulation 434
 ENDOR, electron–nuclear double resonance 262, 274, 276
 energy bandgap 173
 energy bands of silicon 2
 energy conversion efficiency 245
 energy gap 478
 energy loss 216
 energy position of the defect 174
 energy quantization 377
 energy structure 174
 energy transfer 213
 entanglement 467
 enthalpy 173
 entropy 173, 209
 epitaxial deposition processes 44
 epitaxial growth 95
 epitaxial overgrowth 454
 epitaxy 207
 EPD, etch pit density 447

EPM, empirical pseudopotential
method 4

EPR, electron paramagnetic resonance
172, 262

equilibrium critical thickness 100

equilibrium phase diagram 433

error function complements 209

error functions 209

ESR measurements 186

etch-back 153

etch stop(s) 153

eutectic partial systems 433

EUV, extreme ultraviolet 399

evaporation 51

exchange coupling 474

exchange interaction 473

exchange interaction time integral 474

excimer laser processing 61

excitonic effects 6

extended defects 219, 249, 256

extended states 127, 132

external gettering 84

extracellular field potential 528

extrinsic defects 449

extrinsic states 127

Extrovert Gadgets 496

FD, fully depleted SOI process 141

FEC, fully-encapsulated Czochralski
methods 426

Fe silicide precipitates 250

feedback 352

Fermi
energy 264
level 64, 129, 263, 280
level position 249
momentum 475
resonance 272

Feynman 467

FIB, focussed ion beam 386

Fibre Computing 503

Fick's first law 208

field gradient 300

field stop layer 333

figure of merit 353

finger structure 318

FIPOS, full isolation by porous silicon
160

first-order kinetics 267

flat 453

flat-band voltage 417

flex substrate 502

flip chip techniques 501

floating gates 49

floating Si body 383

Fluorescence Interference Contrast
Microscopy 518

flyback converter 341

flyback converter topology 358

formation energy 268

formation enthalpy 209, 210

forward blocking state 309

forward voltage drop 303

Fourier
transform 4
transform spectrometers 186

FPAS, Fourier photoadmittance
spectroscopy 188

free-carrier concentration 269

free-carrier lifetime 343

free energy 125

free-standing active device circuits
509

Frenkel pair 219

frequency-modulation atomic force
microscopy 196

FRS, forward recoil scattering 157

FS-IGBT, field stop IGBT 344

FTIR, Fourier transform infrared
spectroscopy 188, 447

fullerenes 477

functional potassium channels 521

functional sensing species 419

functional silicon fibre 505

fundamental limits 377

fuzzy matching techniques 491

FZ, float-zone 43, 73

Galvani, Luigi 515

GAS, gadget-ware architectural style
496

gas-phase doping 231, 240

gas source MBE 98

gate current 328

gate voltage 483

Gaussian
concentration distribution 216
doping profile 506

- electron beam 406
- profile 218
- GBs, grain boundaries 62
- GCT, gate-commutated thyristor 317
- generation rate 302
- generation/recombination centers 219
- genetic modifications of the membrane 519
- gettering 44, 64, 254, 437
- gettering processes 83
- g*-factor 469
- g*-factor adjustment 474
- g*-factor engineering 465
- Gibbs free energy 173, 438
- glide plane 100
- glide systems 448
- graded SiGe buffer layer 380
- grain boundary potential barrier 62
- grain boundary recombination 65
- grain structure 51
- grinding 454
- ground-state properties 13, 19
- g*-tensor 469
- GTO, bipolar transistor 296
- GTO, gate turn-off thyristor 294
- H_2^* centre 268
- hairpin dislocations 221
- Hall
 - effect 129, 424
 - measurements 27
 - mobility 449, 452
 - sensors 419
- Hamiltonian 2
- hammerheads 401
- handle wafer 153, 156
- harmonic oscillator 197
- Hartree approximation 3
- HBT, hetero bipolar transistor 424, 430
- heat dissipation 404
- heavy-water-moderated reactors 238
- Hebbian learning 527
- Hellman–Feynman theorem 14
- HEMT, high electron mobility transistor 424, 430
- Hertz-type loading 452
- heteroepitaxy 456
- heterogeneous nucleation 56, 441
- heterojunctions 424
- heteropolarity 427
- hexagonal (graphitic) structure 477
- hexagonal wurtzite structure 427
- HGF, horizontal gradient freeze 439
- high-*k* dielectric 385
- high-*k* dielectrics 110
- high-frequency rectifiers 28
- high-*k* materials 38
- high-level lifetime 302, 304, 313
- high-power devices 231
- high-resistivity tunnel junctions 392
- high-voltage switch 342
- Hilbert space 467
- hole mobility 142
- homogeneity 451
- homogeneity range 433
- hopping transport 124, 129
- horizontal-zone refining 78
- hot-wall 434
- HRCs, hydrogen-related cavities 155
- hSlo potassium channels 521
- HVDC, high-voltage direct-current transmission 315
- HWCVD, hot wire chemical vapor deposition 52, 59, 125
- hybrid molecular-Si-CMOS memory array 394
- hybridization 491
- hybrid orbitals 193
- hydrogen bridge bonds 151
- hydrogen-decorated vacancies 155
- hydrogen passivation 86
- hyperfine interaction 265
- hyperfine interactions 279
- hyperfine splitting 279
- hysteresis effects 383
- ICP
 - AES, inductively coupled plasma atomic emission spectrometry 447
 - MS, inductively coupled plasma mass spectrometry 447
- ICs, integrated circuits 49
- ID, inner-diameter sawing 454
- IGBT, insulated-gate bipolar transistor 294, 296, 342
- image distortion 404

impact-ionization generation 351
 Imprint 410
 impurity complex 277
 impurity segregation 62
 impurity striation 232
 IMUs, inertial measurement unit 492
 incoherent lamp irradiation 222
 induced-dipole model 417
 induction heating 77
 inexhaustible sink 221
 infrared detectors 235
 infrared radiation 314
 injection of electrons 107
 inking 411
 in situ cleaning 99
 integrated circuits 36
 intelligent seed programme 497
 interconnect delays 376
 interdigitated finger structure 314
 interface state density distribution D_{it} 219
 interface state 37, 66
 interfaces 246
 interference 417
 internal electric field 80, 254
 internal oxidation process 149
 internal stress 76
 interstitial impurity 261
 interstitial mechanism 210
 interstitial site 247
 interstitial solubility 249
 interstitialcy mechanism 210
 intersubband transition 105, 112
 intervalley scattering 379
 intracellular cytoplasm 515
 intracellular depolarization 525
 intracellular potential 524
 intracellular presynaptic stimulation 527
 intragrain defect 82, 83
 intrinsic acceptor 449
 intrinsic conduction 27
 intrinsic defect 209
 intrinsic electrical conductivity 207
 intrinsic gettering 255
 intrinsic interstitial diffusion coefficient 248
 intrinsic switching speed T_{dn} 377
 inversion layer 321, 327
 ion channel 515
 ion-core potential 9
 ion exchanger 417
 ionic bonding 261
 ionic signals 516
 ion implantation 44, 207
 ionised-impurity scattering 269
 iono-electronic devices 515
 ion selectivity 416
 ion sensitivity mechanism 417
 iron–boron pairs 257
 IR sensors 418
 irreversible poisoning 416
 Irvin curve 235
 ISFETs, ion-sensitive FETs 415
 isotopes
 non-radioactive 234
 isotopic doping 186
 ITO 134
 ITOX, internal oxidation 148
 ITRS, international technology
 roadmap for semiconductors 363, 375, 477, 484
 Jahn–Teller effect 215
 Jahn–Teller lattice distortions 264
 JFET 345
 Joule losses 354
 JSCTs, junction space charge
 techniques 171, 172
 K⁺ sensors 416
 Karman theory 452
 Keck apparatus 425
 kick-out mechanism 248
 Kirk limit 360
 Klitzing resistance 393
 knock-on effects 105
 knock-on particle 219
 k^*p perturbation theory 465
 Krömer 424
 k -selection rule 128
 k -vector mismatch 484
 laminin 518
 lapping 454
 laser diode 429
 laser plasma source 404
 laser recrystallization 125

lateral epitaxial growth 61
 lateral outdiffusion 221
 lateral scattering 217
 lateral spread 212, 217
 lattice defects 28
 lattice mismatch 99, 142, 380
 lattice-mismatched heterostructures 100
 lattice-parameter calculations 13
 lattice relaxation 172
 lattice thermal conductivity 28
 leakage current 301, 302, 304
 LEC, liquid-encapsulated Czochralski technology 426, 441
 LEDs, light-emitting diodes 429
 LEED, low energy electron diffraction 37, 191
 LEEPL, low-energy electron projection lithography 404, 408
 LEEPL technique 409
 length constant 529
 LIC, laser crystallization 59
 lifetime doping 304, 331
 lifetime of the electrical carriers 219
 light absorption 246
 line splittings 268
 line vector 100
 lipid bilayer 515
 liquid state 15
 load current 342
 local breakdown 223
 local density approximation 9
 local-mode phonon 272
 local nucleation 104
 local vibrational mode spectroscopy 268
 local vibrational mode 271
 localized electronic state 246
 localized energy states 251
 LOCOS, local oxidation of silicon process 219
 Loeffler effect 403, 406
 long-range Coulomb interaction 253
 long-range electrostatic interaction 250
 long-range order 123
 Lorentz force 406
 low-field mobility 379
 low-voltage MOS transistor 324
 LPCVD, low-pressure chemical vapor deposition 52
 LPDs, light point defects 455
 LSS theory 217
 LST (laser scattering tomography) 447
 LVM, local vibrational mode 262
 LVM, local-vibrational-mode spectroscopy 186
 spectroscopy 274
 magnetic levitation system 426
 magnetic-field orientation 470
 masking layer 399
 mass transport 54
 MAS NMR, magic-angle spinning nuclear magnetic resonance 220
 maximum blocking voltage 301
 MBE, molecular beam epitaxy 51, 227, 431
 mc-Si, multicrystalline silicon 73, 82
 MCMs, multichip modules 499
 MCUs, microcomputer units 373
 mean projected range $R_p(x)$ 216
 mechanical-chemical polish 159
 mechano-chemical polishing 455
 membrane conductance 519
 MEMs, microelectromechanical systems 365, 367, 492
 MESFET (metal-semiconductor field-effect-transistor) 430
 metal-IC, metal-induced crystallization 51
 metal impurity 245
 metal silicide 60, 249
 metastable state 125
 MIC, metal-induced crystallization 59
 micro–nano hybrid systems 495
 micro-shadow mask technique 115
 microtwins 142
 microwave 27
 mid-bandgap center 343
 migration enthalpy 210, 247
 Miller voltage 326, 335
 milliped 412
 minority carrier lifetime 50, 237, 257
 minority lifetime engineering 257
 misfit dislocations 100

misfit parameter 211
 ML2, maskless lithography 409
 MLD, modified low-dose SIMOX 148
 mobility 66, 303, 379
 mobility edges 127
 mobility of the electrons 301, 323
 MOCVD 431
 modulated reflectivity spectra 6
 Mössbauer
 experiments 274
 molecular approach 394
 Moore's law 39, 375, 491
 more-than-square 350
 MOS, metal-oxide-semiconductor 139, 294
 MOSFET 375, 376
 MOS transistor 296
 mould 76
 MPUs, microprocessors units 372
 μ c-Si:H, microcrystalline silicon films 126, 135
 multilayer gate dielectric 389
 multilayer interconnects 49
 multiple epitaxy 347
 multiple quantum wells 112
 multiple trapping 274
 multiple zone refining 43
 multiple zone-pulling 33
 multiwalled nanotubes 480
 multi-wire saws 76
 muon 265

NAD⁺ (nicotinamide adenine dinucleotide) 418
 NAND bipolar transistors gates 363
 nanoelectronics 490
 nanotube grids 486
 $N(E)$ density distribution 126
 negative charge state 263
 negative correlation energy U 264
 Nernstian sensitivity 416
 nerve-based ionic processor 515
 network 490
 neural network 495
 neurites 527
 neurobiology 515
 neuronal activity 516, 521, 526
 neuronal excitation 528
 neuroprostheses 515

neutral charge state 265
 neutron activation 33
 neutron flux density 234
 neutron irradiation 233
 neutron transmutation 45, 207
 neutron transmutation doped 311
 neutron-irradiated silicon 232
 NGLs, next-generation lithography 399
 nitrided gate oxide 385
 nitride trapping layer 389
 NMR 126
 non-radiative recombination 124, 132, 431
 non-stationary convection 438
 non-stationary nucleation theory 435
 nonvolatile memory elements 394
 NPT, nonpunch-through 342
 NPT-IGBT, nonpunch-through concept 331
 NRA, nuclear reaction analysis 218
 NTD, neutron transmutation doping 231
 nuclear magnetic moments 263
 nucleation 96
 nucleation phase 56
 numerical aperture 401

off-axis illumination 401
 one-dimensional electrostatics 483
 one-electron approximation 3
 on-state resistance 323
 OPC, optical-proximity-effect corrections 401
 open-base condition 319
 open-circuit voltage 65, 68
 operating temperature 1
 optical beam-deflection 198
 optical binding energy 181, 188
 optical emission rate 182
 optical excitation 19
 optical gap 19
 optical lithography 399
 optical phonon 67
 optoelectronic device 424, 429
 oscillation period 469
 oscillator strength 6
 Ostwald ripening 156, 158, 435
 Ostwald ripening mechanism 157

output capacitance 352
 output characteristics 484
 overlap of the wave functions (entanglement) 466
 overlay accuracy 400
 oxide layer 37
 oxide precipitate 144

P.A.M. Dirac 3
 package-level stacking 501
 parallel cantilevers 412
 paramagnetic resonance 469
 partially depleted SOI 383
 PAS, photoadmittance spectroscopy 189
 passivated complexes 262
 passivation 36, 269, 273
 Pauli matrix 469
 P_b -center 219
 pBN, pyrolytic boron nitride 440
 PC, photoconductivity 130
 p-column 351
 PCT (point contact current) 447
 PD, partially depleted 140
 PDAs, personal digital assistants 493
 PDBFET, "planar-doped-barrier MOSFET" 110
 PECVD, plasma enhanced chemical vapor deposition 52, 58, 125
 periodic table 3
 PERL, passivated emitter rear locally-diffused 87
 perturbed-angular-correlation spectroscopy 272
 Petroff-Kimerling mechanism 437
 PFC, power factor correction 358
 pH measurements 415
 pH sensitivity 417
 phase masks
 ALT, alternating 401
 ATN, attenuated 401
 CRL, chromiumless 401
 phonon frequency spectrum 186
 phonon-assisted 471
 phosphorus donor 237
 photocapacitance technique 182
 photochemical reactions 223
 photoconductivity 124, 171
 photocurrent 172
 photocurrent measurements 182
 photodegradation 133
 photoionisation cross-section 181
 photoluminescence 113, 269, 419
 photon-induced valency change 186
 π -electrons 477
 piezo-spectroscopy 172
 piezoelectric polarization 427
 piezoresistive effect 198
 piezospectroscopic tensor 267
 pileup 351
 pin cells 134
 pixel detectors 430
 PL, photoluminescence 130
 planar channeling 216
 planar electrical core-coat conductor 517
 planar techniques 212
 plasma activation 58
 plasma concentration 356
 plasma purification 74
 plasma torch melting 78
 PN junction capacitance 139
 point defect density 105
 point defect 83, 209
 Poisson
 equation 105, 299
 law 344, 348
 polar fluid 515
 polarization 416
 poly-silicon 376
 poly-silicon spacer gate 381
 polyimide inter-dielectric layer 509
 polyimide substrates 502
 polymer bonding technique 159
 polysilicon gate 68
 Poole-Frenkel effect 252
 porous silicon 45, 159, 419
 porous structure 126
 positron annihilation spectroscopy 105
 postsynaptic signals 527
 potential fluctuations 478
 power conversion 341
 power dissipation 376, 465
 power MOS transistors 294
 power MOSFET 344
 power rectifier 35

precipitates 83, 84, 86
 predominance area diagram 438
 preexponential factor 56, 208
 preferential diffusion of dopants 64
 preferred orientation 57
 pressure transducers 418
 PREVAIL, projection electron variable-axis immersion lens 404, 408
 projected range 144, 149, 155
 projection optics 400
 projection printing 43
 proton implantation 262
 proximal-probe lithographies 411
 proximity effect 407
 PRTP, pulsed rapid thermal processing 60
 pseudo-substrate 109
 pseudomorphic structure 101
 pseudopotential 9
 pseudopotential approximation 3
 PT, punch-through 342
 PT-IGBT, punch-through concept 331
 PTIS, photo-thermal ionisation spectroscopy 172, 182
 pulling techniques 441
 pulsed laser processing 222
 punch-through 310, 312, 331, 333
 purification 27
 PVD, physical vapor deposition 52
 PVs, photovoltaics 73
 QC, quantum cascade 112, 465
 quantization of the energy states 391
 quantum algorithms 467
 quantum coherence 468
 quantum dot 19, 389
 quantum efficiency 132
 quantum mechanical effect 377, 390
 quantum parallelism phenomenon 467
 quantum resistance 481
 quantum well 378
 quasi-atomic control 466
 quasi-crystalline wire-like structure 480
 quasi-quantum wires 475
 qubit, quantum memory unit 465
 quench 17
 quenching 86
 RAD, ribbon against drop pulling process ribbon 79
 RADFETs, Radiation-sensing FETs 418
 radiation damage 44, 207, 262
 radiation damage concentration profile 213
 radiation hardness 146
 radiative tunneling 130
 radioactivity 234, 236
 radiotracer analysis 33
 random direction 216
 random network structure 127
 rapid thermal annealing 222
 Rashba effect 474
 Rashba QC 474
 rate of migration 54
 rate window 179
 Rayleigh criterion 401
 RBS, Rutherford back scattering 217
 RDSon 345
 RDSon*A 346
 rechanneling effect 216
 recombination 252, 302
 recombination center 253, 277, 304
 recombination lifetime 132
 recombination rate 246
 recombination velocity 64
 reconfigurable system-on-chip solutions 492
 reconstruction 11
 recrystallization 220
 recrystallization front 225
 rectifier effect 26
 rectifier 298
 redox equilibria 437
 reflection optic 404
 regenerative process 310
 relaxation 273
 relaxation-type gettering 255
 REM, reflection electron microscopy micrograph 226
 repulsive interaction 116
 residual defect 221
 residual gas 214
 residual impurity 436
 resist 399

resistivity 63
 resistivity striations 233
 resonant tunneling effect 105
 RESURF, reduced surface field 113, 346
 retention time 389
 reticle 400, 401
 reverse blocking state 309
 reverse recovery charge 355
 reverse recovery current 307, 355
 reverse recovery process 305
 reverse recovery time 355
 RF, radio frequency 365
 ring mechanism 210
 ring oscillator 483, 507
 rms displacement 15
 robotics 495
 RTCVD, rapid thermal CVD 53
 RTP, rapid thermal processing 85
 RTR, ribbon-to-ribbon crystal growth 79
 Rutherford
 backscattering 209
 scattering 217
 saddle point 271
 SAM, serviceable available market 367
 scaling 375
 scaling effect 494
 SCALPEL, scattering with angular limitation projection electron beam lithography 404
 scanned electron beam 225
 scanning force 411
 scanning near-field 411
 scanning probe method 411
 scanning tunneling 411
 scattering 212
 scattering processes 407
 scattering-free current transport 487
 scattering-free transport of electron 480
 Schmid factor 448
 Schmitt trigger 525
 Schottky barrier 252
 Schottky barrier diode 416
 Schottky defect 210
 Schrödinger equation 2, 105, 467
 sc-Si
 monocrystalline wafers 54
 single-crystal silicon 73
 S-CSP, chip-scale package 499
 screen-printing technology 88
 second breakdown effects 320, 351, 356
 secondary electron 407
 secondary emission 213
 secondary radiation defect 219
 seed crystal 425
 seed wafer 153, 156
 segregation 84, 211, 256, 451
 segregation coefficient 33, 102, 221
 segregation-type gettering 255
 selective emitters 88
 self-annealing 145
 self-assembling way 395
 self-diffusion 210
 self-diffusion constant 15
 self-interstitial injection 249
 self-interstitials 210, 219, 256
 self-nucleation 54
 self-organized growth 114
 self-sensing cantilever 198
 semiconductor laser diode 424
 sensing element 415
 sensor 50
 sensor architecture 498
 serifs 401
 SET, single electron transistor 394
 Seizinghaus-type gas sensor 417
 SFIL, step and flash imprint lithography 411
 SFM, scanning force microscopy 412
 shallow substitutional donors 261
 shaped-beam 407
 Shockley 293
 short-channel effects 387
 short-circuit capability 320, 336
 Si cap layer 109
 Si doping problem 102
 Si-AA9 265
 SiC precipitate 34
 SiC Schottky diode 359
 Si-db, Si dangling bond 123
 Si-db, unsaturated bond 127
 side reaction 235

sidewall transistor 381
 SiGe heterostructure 378
 SiGe layer 378
 silicidation 376
 silicon antibonding position 280
 silicon carbide 308
 silicon fibre computing 499
 silicon ion implantation 143
 silicon limit 345, 350
 silicon ribbon 74
 silicon surface 37
 silicon transistor 1
 silicon-on-insulator 383
 SIMOX, Separation by IMplanted
 OXygen 143
 SIMS, secondary ion mass spectrometry 217
 Si_3N_4 215
 Si_3N_4 passivation 38
 Si-NL8 centre 276
 Si-NL10(H) centre 276
 single-electron device 392
 single-electron limit 465
 single-nanotube device 485
 single-transient capacitance technique 178
 single-walled 480
 SiP, system-in-a-package 499
 Sirtl etch 211
 skewed Gaussian profile 144
 SLS, sequential lateral solidification 61
 smart artefact 494
 Smart Dust project 492
 smart material 490
 smart polymer 491
 smart surface 490
 Smart-Cut[®] 153, 157
 SMC, silicide-mediated crystallization 60
 SMPS, switch mode power supplies 341
 snap-off behavior 307
 snubber circuit 317
 SoC chip 367
 SoC, system on chip 365, 499
 soft error 140
 soft x-ray lithography 403
 softness factor 355
 SOI, silicon-on-insulator 139, 375
 SOI, silicon-on-insulator material 503
 solar cell 430
 solid ion lattice 515
 solid phase epitaxy 105
 solid state condition 222
 solid-state diffusion 451
 solid-state epitaxial regrowth 219
 solubility 245, 433
 solubility limit 214, 218
 source-drain tunneling leakage current 465
 space charge 300, 306
 space charge density 300
 space charge region 318, 326
 space charge zone 300
 spacer 376
 SPC, solid phase crystallization 51, 59, 125
 SPEAR, solid-phase epitaxy and
 regrowth 142
 species-selective membrane 415
 specific resistance 345
 spherical aberration 406
 spherical integrated circuit 493
 spin density 280
 spin-density distribution 274
 spin-density functional theory 266
 spin-orbit coupling interaction 470
 spin-orbit interaction 471
 spin-orbit splitting 470
 spin orientation 468
 spin resonance frequency 469
 spin-splitting engineering 475
 spin state 473
 splitting 159
 sputtering 51, 218
 SRAMs 140
 stacked layer 346
 stacked tandem structure 135
 stacking fault 142, 219, 256
 stacking fault energy 439
 Staebler-Wronski effect 133
 standard deviation ΔR_p 218
 standing modes of the electromagnetic
 field 517
 STD(H), shallow thermal donor 276

stencil mask 403
 step-and-repeat printing 43
 Step growth 96
 stimulation area 524
 STM, scanning tunneling microscope 191, 411
 stopping power 216
 stored charge 302, 304
 strain 100, 470
 strained-layer epitaxy 99
 strained SiGe 377
 strained silicon 379
 strain energy 116
 Stranski-Krastanov mode 100, 116
 stress analyser 448
 stress-induced valley splitting 108
 stress level 445
 stress relaxation 446
 stretch frequency 268
 stretch mode vibration 272
 structural disorder 123
 subband 478
 submicron integrated circuit 245
 suboxide 425
 substitutional site 207
 subthreshold depolarization 527
 subthreshold slope 394
 subthreshold swing 140
 supercell 9
 superconductor 13, 15
 superjunction device 346
 superposition of phonons 209
 supersaturation 96, 255, 435
 surface damage 45, 58
 surface diffusion 95
 surface free-energy 115
 surface phases 117
 surface pores 160
 surface roughness 159
 surface scattering 379
 surface segregation 99, 102
 surface sputtering 212
 surface state 10, 36
 surface step 95
 sweeping-out process 305
 switching loss 344, 353
 switching speed 465
 synapse 522, 527
 synaptic inward current 530
 synchrotron radiation 403
 systems intelligence 491
 tail current 334, 356
 tail state 132
 tailored band structure 429
 Tauc plot 128
 Taylor series 176
 TCO, transparent conductive oxide 82, 133
 TCS, trichlorosilane 54
 TDCM (time-dependent charge measurement) topography 447
 TDDs, thermal double donors 276
 TDLDA, time-dependent local density approximation 19
 TED, transient enhanced diffusion 102
 TEM cross sections 146
 TEM, transmission electron microscope 225
 temperature coefficient 305
 temperature fluctuation 439
 temperature rise 213
 tensile strain 379, 380
 tensile stress 110
 TEOS, tetra-ethyl-ortho-silicate 386
 TEP, transductive extracellular potential 517
 tetrahedral rotor 272
 TFE, thermal field emission 405
 TFTs, thin film transistors 49
 thermal activation energy 52, 175
 thermal annealing 262
 thermal capture cross-section 178
 thermal conductivity 478
 thermal emission 174
 thermo-mechanical stress 223
 thermodynamic equilibrium condition 45
 thermoplastic relaxation 429, 438
 thin-film solar cell 50
 Thomas-Fermi potential 217
 three-band k^*p perturbation theory 469
 three-terminal device 296
 threshold voltage 322, 326, 376, 384
 thyristor 231, 293, 296

tight-binding method 391
 tight-binding model 215
 tilt type boundaries 63
 tilted sidewall implantation 348
 time-of-flight experiment 129
 TM-H₄ complex 278
 touch-polish step 154
 transconductance 483, 484
 transductive extracellular potential 521, 523
 transfer characteristic 393
 transient capacitance technique 175
 transient dark capacitance technique 176
 transition metal 245
 transport property 123
 trapping problem 35
 trench etching 348
 trench technology 328, 335
 trenches 49
 trigger current 330
 triggering by light 315
 triple p-i-n structure 80
 t-SiO₂, thermally oxidized wafers 54
 TSOP, thin small-outline package 501
 TTM, time to market 366
 tuning range of the *g*-factor 471
 tunnel MIS diode 417
 tunneling 132
 tunneling current 192
 tunnelling motion 272
 turn-off time 312
 turn-on threshold voltage 431
 twin boundary 439
 two-dimensional imagers 49
 two-dimensional subband 105
 two-step excitation process 184

 ULSI, ultra large-scale integration 245
 ultralarge-scale integrated devices 256
 ultrametastability 102
 uniaxial stress 267
 unipolar transistor 26
 unit cell 193
 unitary transformation 468
 unobtrusive hardware 490
 Urbach edge 128
 UTSSi, ultra-thin silicon 143

 vacancy cluster 155
 vacancy-interstitial recombination 210
 VAIL, variable-axis immersion lens 408
 van Arkel method 30
 van der Merwe growth 97
 van der Waals force 150
 VAPs, valence alternation pairs 220
 VB, vertical Bridgman technique 426, 441
 VCz, vapour-pressure-controlled Czochralski 426
 Vegard's law 99
 Vegard's rule 429
 vertical Bridgman 441
 vertical gradient freeze method 441
 vertical transistor 381
 vertical zone melting 30
 VGF 441
 via 481
 via-holes 500
 vibrational frequency 186
 Volmer-Weber island growth 100
 voltage-controlled device 322
 voltage drop 302
 voltage-sensitive ion channel 517
 voltage spike 307, 334

 wafer bonding 150
 wafer-level packaging 500
 wafer tracking 453
 wagging mode 272
 warping 43, 454
 wave function 391
 wavelength 43
 wave vector 2
 wearable computing 503
 Weibull distribution 452
 Welker 424
 WICP, Wacker ingot casting process 75
 wire sawing 454
 work function 11

 X-ray rocking-curve 448
 XPS, x-ray photoemission spectroscopy 7

- Zeeman
 - effect 279
 - frequency 276
 - spectroscopy 172
- Zener effect 301
- Zener tunneling currents 382, 383
- zero-field conditions 180
- zero-field enthalpies 181
- zigzag configuration 478
- ZMR, zone-melting recrystallization
 - 52, 59, 141
- zone melting 232
- ZVS bridge applications 356
- ZVS, zero-voltage-switching 354